SUPERSYMMETRY AND STRING THEORY

Contents

Preface *page* xv
A note on choice of metric xviii
Text website xx
Part 1 Effective field theory: the Standard Model, supersymmetry, unification 1
1 Before the Standard Model 3
Suggested reading 7
2 The Standard Model 9
2.1 Yang–Mills theory 9
2.2 Realizations of symmetry in quantum field theory 12
2.3 The quantization of Yang–Mills theories 18
2.4 The particles and fields of the Standard Model 22
2.5 The gauge boson masses 25
2.6 Quark and lepton masses 27
Suggested reading 28
Exercises 28
3 Phenomenology of the Standard Model 29
3.1 The weak interactions 29
3.2 The quark and lepton mass matrices 32
3.3 The strong interactions 34
3.4 The renormalization group 35
3.5 Calculating the beta function 39
3.6 The strong interactions and dimensional transmutation 43
3.7 Confinement and lattice gauge theory 44
3.8 Strong interaction processes at high momentum transfer 51
Suggested reading 59
Exercises 61

4 The Standard Model as an effective field theory 63
4.1 Lepton and baryon number violation 66
4.2 Challenges for the Standard Model 70
4.3 The hierarchy problem 71
4.4 Dark matter and dark energy 72
4.5 Summary: successes and limitations of the Standard Model 73
Suggested reading 73
5 Anomalies, instantons and the strong CP problem 75
5.1 The chiral anomaly 76
5.2 A two-dimensional detour 81
5.3 Real QCD 89
5.4 The strong CP problem 100
5.5 Possible solutions of the strong CP problem 102
Suggested reading 105
Exercises 106
6 Grand unification 107
6.1 Cancellation of anomalies 110
6.2 Renormalization of couplings 110
6.3 Breaking to $SU(3) \times SU(2) \times U(1)$ 111
6.4 $SU(2) \times U(1)$ breaking 112
6.5 Charge quantization and magnetic monopoles 113
6.6 Proton decay 114
6.7 Other groups 114
Suggested reading 117
Exercises 117
7 Magnetic monopoles and solitons 119
7.1 Solitons in $1 + 1$ dimensions 120
7.2 Solitons in $2 + 1$ dimensions: strings or vortices 122
7.3 Magnetic monopoles 122
7.4 The BPS limit 124
7.5 Collective coordinates for the monopole solution 125
7.6 The Witten effect: the electric charge in the presence of θ 127
7.7 Electric–magnetic duality 128
Suggested reading 129
Exercises 129
8 Technicolor: a first attempt to explain hierarchies 131
8.1 QCD in a world without Higgs fields 132
8.2 Fermion masses: extended technicolor 133

8.3 Precision electroweak measurements 135
Suggested reading 136
Exercises 136
Part 2 Supersymmetry 137
9 Supersymmetry 139
9.1 The supersymmetry algebra and its representations 140
9.2 Superspace 140
9.3 $N = 1$ Lagrangians 144
9.4 The supersymmetry currents 147
9.5 The ground-state energy in globally supersymmetric theories 148
9.6 Some simple models 149
9.7 Non-renormalization theorems 151
9.8 Local supersymmetry: supergravity 154
Suggested reading 155
Exercises 155
10 A first look at supersymmetry breaking 157
10.1 Spontaneous supersymmetry breaking 157
10.2 The goldstino theorem 160
10.3 Loop corrections and the vacuum degeneracy 161
10.4 Explicit, soft supersymmetry breaking 162
10.5 Supersymmetry breaking in supergravity models 163
Suggested reading 166
Exercises 166
11 The Minimal Supersymmetric Standard Model 167
11.1 Soft supersymmetry breaking in the MSSM 169
11.2 $SU(2) \times U(1)$ breaking 173
11.3 Why is one Higgs mass negative? 175
11.4 Radiative corrections to the Higgs mass limit 176
11.5 Embedding the MSSM in supergravity 177
11.6 The μ term 178
11.7 Constraints on soft breakings 179
Suggested reading 183
Exercises 183
12 Supersymmetric grand unification 185
12.1 A supersymmetric grand unified model 185
12.2 Coupling constant unification 186
12.3 Dimension-five operators and proton decay 188
Suggested reading 189
Exercises 189

13 Supersymmetric dynamics 191
13.1 Criteria for supersymmetry breaking: the Witten index 192
13.2 Gaugino condensation in pure gauge theories 193
13.3 Supersymmetric QCD 194
13.4 $N_f < N$: a non-perturbative superpotential 197
13.5 The superpotential in the case $N_f < N - 1$ 200
13.6 $N_f = N - 1$: the instanton-generated superpotential 201
Suggested reading 208
Exercises 208
14 Dynamical supersymmetry breaking 209
14.1 Models of dynamical supersymmetry breaking 209
14.2 Particle physics and dynamical supersymmetry breaking 211
Suggested reading 218
Exercises 218
15 Theories with more than four conserved supercharges 219
15.1 $N = 2$ theories: exact moduli spaces 219
15.2 A still simpler theory: $N = 4$ Yang–Mills 221
15.3 A deeper understanding of the BPS condition 223
15.4 Seiberg–Witten theory 225
Suggested reading 230
Exercises 231
16 More supersymmetric dynamics 233
16.1 Conformally invariant field theories 233
16.2 More supersymmetric QCD 235
16.3 $N_f = N_c$ 236
16.4 $N_f > N + 1$ 240
16.5 $N_f \geq 3/2N$ 241
Suggested reading 241
Exercises 242
17 An introduction to general relativity 243
17.1 Tensors in general relativity 244
17.2 Curvature 249
17.3 The gravitational action 250
17.4 The Schwarzschild solution 252
17.5 Features of the Schwarzschild metric 254
17.6 Coupling spinors to gravity 256
Suggested reading 257
Exercises 257
18 Cosmology 259
18.1 A history of the universe 263

Suggested reading 268
Exercises 268
19 Astroparticle physics and inflation 269
19.1 Inflation 272
19.2 The axion as dark matter 280
19.3 The LSP as the dark matter 283
19.4 The moduli problem 285
19.5 Baryogenesis 287
19.6 Flat directions and baryogenesis 294
19.7 Supersymmetry breaking in the early universe 296
19.8 The fate of the condensate 297
19.9 Dark energy 300
Suggested reading 301
Exercises 301
Part 3 String theory 303
20 Introduction 305
20.1 The peculiar history of string theory 306
Suggested reading 311
21 The bosonic string 313
21.1 The light cone gauge in string theory 315
21.2 Closed strings 318
21.3 String interactions 320
21.4 Conformal invariance 322
21.5 Vertex operators and the S-matrix 328
21.6 The S-matrix vs. the effective action 334
21.7 Loop amplitudes 335
Suggested reading 338
Exercises 338
22 The superstring 341
22.1 Open superstrings 341
22.2 Quantization in the Ramond sector: the appearance of space-time fermions 343
22.3 Type II theory 344
22.4 World sheet supersymmetry 345
22.5 The spectra of the superstrings 346
22.6 Manifest space-time supersymmetry: the Green–Schwarz formalism 353
22.7 Vertex operators 355
Suggested reading 356
Exercises 356

23 The heterotic string 359
23.1 The $O(32)$ theory 360
23.2 The $E_8 \times E_8$ theory 361
23.3 Heterotic string interactions 361
23.4 A non-supersymmetric heterotic string theory 363
Suggested reading 363
Exercises 364
24 Effective actions in ten dimensions 365
24.1 Coupling constants in string theory 368
Suggested reading 371
Exercise 371
25 Compactification of string theory I. Tori and orbifolds 373
25.1 Compactification in field theory: the Kaluza–Klein program 373
25.2 Closed strings on tori 377
25.3 Enhanced symmetries 380
25.4 Strings in background fields 382
25.5 Bosonic formulation of the heterotic string 386
25.6 Orbifolds 387
25.7 Effective actions in four dimensions for orbifold models 395
25.8 Non-supersymmetric compactifications 398
Suggested reading 399
Exercises 400
26 Compactification of string theory II. Calabi–Yau compactifications 401
26.1 Mathematical preliminaries 401
26.2 Calabi–Yau spaces: constructions 406
26.3 The spectrum of Calabi–Yau compactifications 409
26.4 World sheet description of Calabi–Yau compactification 411
26.5 An example: the quintic in CP^4 414
26.6 Calabi–Yau compactification of the heterotic string at weak coupling 416
Suggested reading 426
Exercises 427
27 Dynamics of string theory at weak coupling 429
27.1 Non-renormalization theorems 430
27.2 Fayet–Iliopoulos D-terms 434
27.3 Gaugino condensation 438
27.4 Obstacles to a weakly coupled string phenomenology 439
Suggested reading 440
28 Beyond weak coupling: non-perturbative string theory 441
28.1 Perturbative dualities 442

28.2 Strings at strong coupling: duality 442
28.3 *D*-branes 443
28.4 Branes from *T*-duality of Type I strings 447
28.5 Strong–weak coupling dualities: the equivalence of different string theories 451
28.6 Strong–weak coupling dualities: some evidence 452
28.7 Strongly coupled heterotic string 458
28.8 Non-perturbative formulations of string theory 460
Suggested reading 465
Exercises 466
29 Large and warped extra dimensions 467
29.1 Large extra dimensions: the ADD proposal 467
29.2 Warped spaces: the Randall–Sundrum proposal 470
Suggested reading 473
Exercise 473
30 Coda: where are we headed? 475
Suggested reading 479
Part 4 The appendices 481
Appendix A Two-component spinors 483
Appendix B Goldstone's theorem and the pi mesons 487
Exercises 489
Appendix C Some practice with the path integral in field theory 491
C.1 Path integral review 491
C.2 Finite-temperature field theory 492
C.3 QCD at high temperature 495
C.4 Weak interactions at high temperature 496
C.5 Electroweak baryon number violation 497
Suggested reading 499
Exercises 499
Appendix D The beta function in supersymmetric Yang–Mills theory 501
Exercise 503
References 505
Index 511

Preface

As this is being written, particle physics stands on the threshold of a new era, with the commissioning of the Large Hadron Collider (LHC) not even two years away. In writing this book, I hope to help prepare graduate students and postdoctoral researchers for what will hopefully be a period rich in new data and surprising phenomena.

The Standard Model has reigned triumphant for three decades. For just as long, theorists and experimentalists have speculated about what might lie beyond. Many of these speculations point to a particular energy scale, the teraelectronvolt (TeV) scale which will be probed for the first time at the LHC. The stimulus for these studies arises from the most mysterious – and still missing – piece of the Standard Model: the Higgs boson. Precision electroweak measurements strongly suggest that this particle is elementary (in that any structure is likely far smaller than its Compton wavelength), and that it should be in a mass range where it will be discovered at the LHC. But the existence of fundamental scalars is puzzling in quantum field theory, and strongly suggests new physics at the TeV scale. Among the most prominent proposals for this physics is a hypothetical new symmetry of nature, supersymmetry, which is the focus of much of this text. Others, such as technicolor, and large or warped extra dimensions, are also treated here.

Even as they await evidence for such new phenomena, physicists have become more ambitious, attacking fundamental problems of quantum gravity, and speculating on possible final formulations of the laws of nature. This ambition has been fueled by *string theory*, which seems to provide a complete framework for the quantum mechanics of gauge theory and gravity. Such a structure is necessary to give a framework to many speculations about beyond the Standard Model physics. Most models of supersymmetry breaking, theories of large extra dimensions, and warped spaces cannot be discussed in a consistent way otherwise.

It seems, then, quite likely that a twentyfirst-century particle physicist will require a working knowledge of supersymmetry and string theory, and in writing this

text I hope to provide this. The first part of the text is a review of the Standard Model. It is meant to complement existing books, providing an introduction to perturbative and phenomenological aspects of the theory, but with a lengthy introduction to non-perturbative issues, especially in the strong interactions. The goal is to provide an understanding of chiral symmetry breaking, anomalies and instantons, suitable for thinking about possible strong dynamics, and about dynamical issues in supersymmetric theories. The first part also introduces grand unification and magnetic monopoles.

The second part of the book focuses on supersymmetry. In addition to global supersymmetry in superspace, there is a study of the supersymmetry currents, which are important for understanding dynamics, and also for understanding the BPS conditions which play an important role in field theory and string theory dualities. The MSSM is developed in detail, as well as the basics of supergravity and supersymmetry breaking. Several chapters deal with supersymmetry dynamics, including dynamical supersymmetry breaking, Seiberg dualities and Seiberg–Witten theory. The goal is to introduce phenomenological issues (such as dynamical supersymmetry breaking in hidden sectors and its possible consequences), and also to illustrate the control that supersymmetry provides over dynamics.

I then turn to another critical element of beyond the Standard Model physics: general relativity, cosmology and astrophysics. The chapter on general relativity is meant as a brief primer. The approach is more field theoretic than geometrical, and the uninitiated reader will learn the basics of curvature, the Einstein Lagrangian, the stress tensor and equations of motion, and will encounter the Schwarzschild solution and its features. The subsequent two chapters introduce the basic features of the FRW cosmology, and then very early universe cosmology: cosmic history, inflation, structure formation, dark matter and dark energy. Supersymmetric dark matter and axion dark matter, and mechanisms for baryogenesis, are all considered.

The third part of the book is an introduction to string theory. My hope, here, is to be reasonably comprehensive while not being excessively technical. These chapters introduce the various string theories, and quickly compute their spectra and basic features of their interactions. Heavy use is made of light cone methods. The full machinery of conformal and superconformal ghosts is described but not developed in detail, but conformal field theory techniques are used in the discussion of string interactions. Heavy use is also made of effective field theory techniques, both at weak and strong coupling. Here, the experience in the first half of the text with supersymmetry is invaluable; again supersymmetry provides a powerful tool to constrain and understand the underlying dynamics. Two lengthy chapters deal with string compactifications; one is devoted to toroidal and orbifold compactifications, which are described by essentially free strings; the other introduces the basics of Calabi–Yau compactification. Four appendices make up the final part of this book.

The emphasis in all of this discussion is on providing tools with which to consider how string theory might be related to observed phenomena. The obstacles are made clear, but promising directions are introduced and explored. I also attempt to stress how string theory can be used as a testing ground for theoretical speculations. I have not attempted a complete bibliography. The suggested reading in each chapter directs the reader to a sample of reviews and texts.

What I know in field theory and string theory is the result of many wonderful colleagues. It is impossible to name all of them, but Tom Appelquist, Nima Arkani-Hamed, Tom Banks, Savas Dimopoulos, Willy Fischler, Michael Green, David Gross, Howard Haber, Jeff Harvey, Shamit Kachru, Andre Linde, Lubos Motl, Ann Nelson, Yossi Nir, Michael Peskin, Joe Polchinski, Pierre Ramond, Lisa Randall, John Schwarz, Nathan Seiberg, Eva Silverstein, Bunji Sakita, Steve Shenker, Leonard Susskind, Scott Thomas, Steven Weinberg, Frank Wilczek, Mark Wise and Edward Witten have all profoundly influenced me, and this influence is reflected in this text. Several of them offered comments on the text or provided specific advice and explanations, for which I am grateful. I particularly wish to thank Lubos Motl for reading the entire manuscript and correcting numerous errors. Needless to say, none of them are responsible for the errors which have inevitably crept into this book.

Some of the material, especially on anomalies and aspects of supersymmetry phenomenology, has been adapted from lectures given at the Theoretical Advanced Study Institute, held in Boulder, Colorado. I am grateful to K. T. Manahathapa for his help during these schools, and to World Scientific for allowing me to publish these excerpts. The lectures "Supersymmetry Phenomenology with a Broad Brush" appeared in *Fields, Strings and Duality*, ed. C. Efthimiou and B. Greene (Singapore: World Scientific, 1997); "TASI Lectures on M Theory Phenomenology" appeared in *Strings, Branes and Duality*, ed. C. Efthimiou and B. Greene (Singapore: World Scientific, 2001); and "The Strong CP Problem" in *Flavor Physics for the Millennium: TASI 2000*, ed. J. L. Rosner (Singapore: World Scientific, 2000).

I have used much of the material in this book as the basis for courses, and I am also grateful to students and postdocs (especially Patrick Fox, Assaf Shomer, Sean Echols, Jeff Jones, John Mason, Alex Morisse, Deva O'Neil, and Zheng Sun) at Santa Cruz who have patiently suffered through much of this material as it was developed. They have made important comments on the text and in the lectures, often filling in missing details. As teachers, few of us have the luxury of devoting a full year to topics such as this. My intention is that the separate supersymmetry or string parts are suitable for a one-quarter or one-semester special topics course.

Finally, I wish to thank Aviva, Jeremy, Shifrah, and Melanie for their love and support.

A note on choice of metric

There are two popular choices for the metric of flat Minkowski space. One, often referred to as the "West Coast Metric," is particularly convenient for particle physics applications. Here,

$$ds^2 = dt^2 - d\vec{x}^2 = \eta_{\mu\nu} dx^\mu dx^\nu \tag{0.1}$$

This has the virtue that $p^2 = E^2 - \vec{p}^2 = m^2$. It is the metric of many standard texts in quantum field theory. But it has the annoying feature that ordinary, space-like intervals – conventional lengths – are treated with a minus sign. So in most general relativity textbooks, as well as string theory textbooks, the "East Coast Metric" is standard:

$$ds^2 = -dt^2 + d\vec{x}^2. \tag{0.2}$$

Many physicists, especially theorists, become so wedded to one form or another that they resist – or even have difficulty – switching back and forth. This is a text, however, meant to deal both with particle physics and with general relativity and string theory. So, in the first half of the book, which deals mostly with particle physics and quantum field theory, we will use the "West Coast" convention. In the second half, dealing principally with general relativity and string theory, we will switch to the "East Coast" convention. For both the author and the readers, this may be somewhat disconcerting. While I have endeavored to avoid errors from this somewhat schizophrenic approach, some have surely slipped by. But I believe that this freedom to move back and forth between the two conventions will be both convenient and healthy. If nothing else, this is probably the first textbook in physics in which the author has deliberately used both conventions (many have done so inadvertently).

At a serious level, the researcher must always be careful in computations to be consistent. It is particularly important to be careful in borrowing formulas from

papers and texts, and especially in downloading computer programs, to make sure one has adequate checks on such matters of signs. I will appreciate being informed of any such inconsistencies, as well as of other errors, both serious and minor, which have crept into this text.

Text website

Even as this book was going to press, there were important developments in a number of these subjects. The website http://scipp.ucsc.edu/∼dine/book/book.html will contain

(1) updates,
(2) errata,
(3) solutions of selected problems, and
(4) additional selected reading.

Part 1

Effective field theory: the Standard Model, supersymmetry, unification

1

Before the Standard Model

Two of the most profound scientific discoveries of the early twentieth century were special relativity and quantum mechanics. With special (and general) relativity came the notion that physics should be local. Interactions should be carried by dynamical fields in space-time. Quantum mechanics altered the questions which physicists asked about phenomena; the rules governing microscopic (and some macroscopic) phenomena were not those of classical mechanics. When these ideas are combined, they take on their full force, in the form of *quantum field theory*. Particles themselves are localized, finite-energy excitations of fields. Otherwise mysterious phenomena such as the connection of spin and statistics are immediate consequences of this marriage. But quantum field theory does pose a serious challenge. The Schrödinger equation seems to single out time, making a manifestly relativistic description difficult. More serious, but closely related, the number of degrees of freedom is infinite. In the 1920s and 1930s, physicists performed conventional perturbation theory calculations in the quantum theory of electrodynamics, quantum electrodynamics or QED, and obtained expressions which were neither Lorentz invariant nor finite. Until the late 1940s, these problems stymied any quantitative progress, and there was serious doubt whether quantum field theory was a sensible framework for physics.

Despite these concerns, quantum field theory proved a valuable tool with which to consider problems of fundamental interactions. Yukawa proposed a field theory of the nuclear force, in which the basic quanta were mesons. The corresponding particle was discovered shortly after the Second World War. Fermi was aware of Yukawa's theory, and proposed that the weak interactions arose through the exchange of some massive particle – essentially the $W^{\pm}$ bosons which were finally discovered in the 1980s. The large mass of the particle accounted for both the short range and the strength of the weak force. Because of the very short range of the force, one could describe it in terms of four fields interacting at a point. In the early days of the theory, these were the proton, neutron, electron and neutrino.

Viewed as a theory of four-fermion interactions, Fermi's theory was very successful, accounting for all experimental weak interaction results until well into the 1970s. Yet the theory raised even more severe conceptual problems than QED. At high energies, the amplitudes computed in the leading approximation violated unitarity, and higher-order terms in perturbation theory were very divergent.

The difficulties of QED were overcome in the late 1940s, by Bethe, Dyson, Feynman, Schwinger, Tomanaga and others, as experiments in atomic physics demanded high-precision QED calculations. As a result of their work, it was now possible to perform perturbative calculations in a manifestly Lorentz invariant fashion. Exploiting the covariance, the infinities could be controlled and, over time, their significance came to be understood. Quantum electrodynamics achieved extraordinary successes, explaining the magnetic moment of the electron to extraordinary precision, as well as the Lamb shift in hydrogen and other phenomena. One now, for the first time, had an example of a system of physical law, consistent both with Einstein's principles of relativity and with quantum mechanics.

There were, however, many obstacles to extending this understanding to the strong and weak interactions, and at times it seemed that some other framework might be required. The difficulties came in various types. The infinities of Fermi's theory could not be controlled as in electrodynamics. Even postulating the existence of massive particles to mediate the force did not solve the problems. But the most severe difficulties came in the case of the strong interactions. The 1950s and 1960s witnessed the discovery of hundreds of hadronic resonances. It was hard to imagine that each of these should be described by still another fundamental field. Some theorists pronounced field theory dead, and sought alternative formulations (among the outgrowths of these explorations was string theory, which has emerged as the most promising setting for a quantum theory of gravitation). Gell-Mann and Zweig realized that quarks could serve as an organizing principle. Originally, there were only three, u, d, and s, with baryon number $1/3$ and charges $2/3$, $-1/3$ and $-1/3$ respectively. All of the known hadrons could be understood as bound states of objects with these quantum numbers. Still, there remained difficulties. First, the quarks were strongly interacting, and there were no successful ideas for treating strongly interacting fields. Second, searches for quarks came up empty handed.

In the late 1960s, a dramatic series of experiments at SLAC, and a set of theoretical ideas due to Feynman and Bjorken, changed the situation again. Feynman had argued that one should take seriously the idea of quarks as dynamical entities (for a variety of reasons, he hesitated to call them quarks, referring to them as "partons"). He conjectured that these partons would behave as nearly free particles in situations where momentum transfers were large. He and Bjorken realized that this picture implied scaling in deep inelastic scattering phenomena. The experiments at SLAC

exhibited just this phenomenon and showed that the partons carried the electric charges of the u and d quarks.

But this situation was still puzzling. Known field theories did not behave in the fashion conjectured by Feynman and Bjorken. The interactions of particles typically became *stronger* as the energies and momentum transfers grew. This is the case, for example, in quantum electrodynamics, and a simple quantum mechanical argument, based on unitarity and relativity, would seem to suggest it is general. But there turned out to be an important class of theories with the opposite property.

In 1954, Yang and Mills wrote down a generalization of electrodynamics, where the $U(1)$ symmetry group is enlarged to a non-Abelian group. While mathematically quite beautiful, these *non-Abelian gauge theories* remained oddities for some time. First, their possible place in the scheme of things was not known (Yang and Mills themselves suggested that perhaps their vector particles were the ρ mesons). Moreover, their quantization was significantly more challenging than that of electrodynamics. It was not at all clear that these theories really made sense at the quantum level, respecting both the principles of Lorentz invariance and unitarity. The first serious effort to quantize Yang–Mills theories was probably due to Schwinger, who chose a non-covariant but manifestly unitary gauge, and carefully verified that the Poincaré algebra is satisfied. The non-covariant gauge, however, was exceptionally awkward. Real progress in formulating a covariant perturbation expansion was due to Feynman, who noted that naive Feynman rules for these theories were not unitary, but that this could be fixed, at least in low orders, by adding a set of fictitious fields ("ghosts"). A general formulation was provided by Faddeev and Popov, who derived Feynman's covariant rules in a path integral formulation, and showed their formal equivalence to Schwinger's manifestly unitary formulation. A convincing demonstration that these theories were unitary, covariant *and renormalizable* was finally given in the early 1970s by 't Hooft and Veltman, who developed elegant and powerful techniques for doing real calculations as well as formal proofs.

In the original Yang–Mills theories, the vector bosons were massless and their possible connections to known phenomena obscure. However, Peter Higgs discovered a mechanism by which these particles could become massive. In 1967, Weinberg and Salam wrote down a Yang–Mills theory of the weak interactions based on the Higgs mechanism. This finally realized Fermi's idea that the weak interactions arise from the exchange of a very massive particle. To a large degree this work was ignored, until 't Hooft and Veltman proved the unitarity and renormalizability of these theories. At this point the race to find precisely the correct theory and study its experimental consequences was on; Weinberg and Salam's first guess turned out to be correct.

The possible role of Yang–Mills fields in strong interactions was, at first sight, even more obscure. To complete the story required another important fact of

hadronic physics. While the quark model was very successful, it was also puzzling. The quarks were spin $1/2$ particles, yet models of the hadrons seemed to require that the hadronic wave functions were symmetric under interchange of quark quantum numbers. A possible resolution, suggested by Greenberg, was that the quarks carried an additional quantum number, called color, coming in three possible types. The statistics puzzle was solved if the hadron wave functions were totally antisymmetric in color. This hypothesis required that the color symmetry, unlike, say, isospin, would be exact, and thus special. While seemingly contrived, it explained two other facts: the width of the π^0 meson (which was otherwise too small by a factor of three) and the value of the e^+e^- cross section to hadrons.

To a number of researchers, the exactness of this symmetry suggested a possible role for Yang–Mills theory. So, in retrospect, there was the obvious question: could it be that an $SU(3)$ Yang–Mills theory, describing the interactions of quarks, would exhibit the property required to explain Bjorken scaling: the interactions become weak at short distances? Of course, things were not quite so obvious at the time.The requisite calculation had already been done by 't Hooft, but the result seems not to have been widely known or its significance appreciated. David Gross and his student Frank Wilczek set out to prove that no field theory had the required scaling property, while Sidney Coleman, apparently without any particular prejudice, assigned the problem to his graduate student, David Politzer. All soon realized that Yang–Mills theories do have the property of asymptotic freedom: the interactions become weak at high-momentum transfers or at short distances.

Experiment and theory now entered a period of remarkable convergence. Alternatives to the Weinberg–Salam theory were quickly ruled out. The predictions of QCD were difficult, at first, to verify in detail. The theory predicted small violations of Bjorken scaling, depending logarithmically on energy, and it took many years to convincingly measure them. But there was another critical experimental development which clinched the picture. The existence of a heavy quark beyond the u, d and s, had been predicted by Glashow, Iliopoulos and Maiani, and was a crucial part of the developing Standard Model. The mass of this *charm* quark had been estimated by Gaillard and Lee. Appelquist and Politzer predicted, almost immediately after the discovery of asymptotic freedom, that heavy quarks would be bound in narrow vector resonances. In 1974 a narrow resonance was discovered in e^+e^- annihilation, the J/ψ particle, which was quickly identified as a bound state of a charmed quark and its antiparticle.

Over the next 25 years, this Standard Model was subjected to more and more refined tests. One feature absent from the original Standard Model was CP (T) violation. Kobiyashi and Maskawa pointed out that, if there were a third generation of quarks and leptons, the theory could accommodate the observed CP violation in the K meson system. Two more quarks and a lepton were discovered, and their

interactions and behavior were as expected within the Standard Model. Jets of particles which could be associated with gluons were seen in the late 1970s. The W and Z particles were produced in accelerators in the early 1980s. At CERN and SLAC, precision measurements of the Z mass and width provided stringent tests of the weak-interaction part of the theory. Detailed measurements in deep inelastic scattering and in jets provided precise confirmation of the logarithmic scaling violations predicted by QCD. The Standard Model passed every test.

So why write a book about physics beyond the Standard Model? For all of its simplicity and success in reproducing the interactions of elementary particles, the Standard Model cannot represent a complete description of nature. In the first few chapters of this book, we will review the Standard Model and its successes, and then we will discuss some of the theory's limitations. These include the *hierarchy problem*, which, at its most primitive level, represents a failure of dimensional analysis; the presence of a large number of parameters; the incompatibility of quantum mechanics with Einstein's theory of general relativity; the inability of the theory to account for the small, non-zero value of the cosmological constant (an even more colossal failure of dimensional analysis), and the failure of the theory to account for basic features of our universe: the excess of baryons over anti-baryons, the dark matter, and structure. Then we will set out on an exploration of possible phenomena which might address these questions. These include supersymmetry, technicolor and large or warped extra dimensions, as possible solutions to the hierarchy problem; grand unification, as a partial solution to the overabundance of parameters, and superstring theory, as a possible solution to the problem of quantizing gravity, which incorporates many features of these other proposals.

Suggested reading

A complete bibliography of the Standard Model would require a book by itself. A good deal of the history of special relativity, quantum mechanics and quantum field theory can be found in *Inward Bound*, by Abraham Pais (1986), which also includes an extensive bibliography. The development of the Standard Model is also documented in this very readable book. As a minor historical note, I would add that the earliest reference where I have heard the observation that a Yang–Mills theory might underly the strong interactions is due to Feynman, *c*. 1963 (Roger Dashen, personal communication, 1981), who pointed out that, in an $SU(3)$ Yang–Mills theory, three quarks would be bound together, as well as quark–antiquark pairs.

2

The Standard Model

The interactions of the Standard Model give rise to the phenomena of our day to day experience. They explain virtually all of the particles and interactions which have been observed in accelerators. Yet the underlying laws can be summarized in a few lines. In this chapter, we describe the ingredients of this theory and some of its important features. Many dynamical questions will be studied in subsequent chapters. For detailed comparison of theory and experiment, there are a number of excellent texts, described in the suggested reading at the end of the chapter.

2.1 Yang–Mills theory

By the early 1950s, physicists were familiar with approximate global symmetries like isospin. Yang and Mills argued that the lesson of Einstein's general theory was that symmetries, if fundamental, should be local. In ordinary electrodynamics, the gauge symmetry is a local, Abelian symmetry. Yang and Mills explained how to generalize this to a non-Abelian symmetry group. Let's first review the case of electrodynamics. The electron field, $\psi(x)$, transforms under a gauge transformation as:

$$\psi(x) \to e^{i\alpha(x)}\psi(x) = g_\alpha(x)\psi(x). \tag{2.1}$$

We can think of $g_\alpha(x) = e^{i\alpha(x)}$ as a group element in the group $U(1)$. The group is Abelian, $g_\alpha g_\beta = g_\beta g_\alpha = g_{\alpha+\beta}$. Quantities like $\bar{\psi}\psi$ are gauge invariant, but derivative terms, like $i\bar{\psi}\not{\partial}\psi$, are not. In order to write derivative terms in an action or equations of motion, one needs to introduce a gauge field, A_μ, transforming under the symmetry transformation as:

$$\begin{aligned} A_\mu &\to A_\mu + \partial_\mu \alpha \\ &= A_\mu + ig(x)\partial_\mu g^\dagger(x). \end{aligned} \tag{2.2}$$

This second form allows generalization more immediately to the non-Abelian case. Given A_μ and its transformation properties, we can define a covariant derivative:

$$D_\mu \psi = (\partial_\mu - i A_\mu)\psi. \tag{2.3}$$

This derivative has the property that it transforms like ψ itself under the symmetry:

$$D_\mu \psi \to g(x) D_\mu \psi. \tag{2.4}$$

We can also form a gauge-invariant object out of the gauge fields, A_μ, themselves. A simple way to do this is to construct the commutator of two covariant derivatives:

$$F_{\mu\nu} = i[D_\mu, D_\nu] = \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{2.5}$$

This form of the gauge transformations may be somewhat unfamiliar. Note, in particular, that the charge of the electron, e (the gauge coupling) does not appear in the transformation laws. Instead, the gauge coupling appears when we write a gauge-invariant Lagrangian:

$$\mathcal{L} = i\bar{\psi}\, \not{D}\psi - m\bar{\psi}\psi - \frac{1}{4e^2} F_{\mu\nu}^2. \tag{2.6}$$

The more familiar formulation is obtained if we make the replacement:

$$A_\mu \to e A_\mu. \tag{2.7}$$

In terms of this new field, the gauge transformation law is:

$$A_\mu \to A_\mu + \frac{1}{e}\partial_\mu \alpha \tag{2.8}$$

and the covariant derivative is:

$$D_\mu \psi = (\partial_\mu - ieA_\mu)\psi. \tag{2.9}$$

We can generalize this to a non-Abelian group, $\mathcal{G}$, by taking ψ to be a field (fermion or boson) in some representation of the group. $g(x)$ is then a matrix which describes a group transformation acting in this representation. Formally, the transformation law is the same as before:

$$\psi \to g(x)\psi(x), \tag{2.10}$$

but the group composition law is more complicated,

$$g_\alpha g_\beta \neq g_\beta g_\alpha. \tag{2.11}$$

The gauge field, A_μ is now a matrix-valued field, transforming in the adjoint representation of the gauge group:

$$A_\mu \to g A_\mu g^\dagger + i g(x)\partial_\mu g^\dagger(x). \tag{2.12}$$

The covariant derivative also, formally, looks exactly as before:

$$D_\mu \psi = (\partial_\mu - i A_\mu)\psi; \quad D_\mu \psi \to g(x) D_\mu \psi. \tag{2.13}$$

Like A_μ, the field strength is a matrix-valued field:

$$F_{\mu\nu} = i[D_\mu, D_\nu] = \partial_\mu A_\nu - \partial_\nu A_\mu - i[A_\mu, A_\nu]. \tag{2.14}$$

Note that $F_{\mu\nu}$ is not gauge *invariant*, but rather covariant:

$$F_{\mu\nu} \to g F_{\mu\nu} g^\dagger, \tag{2.15}$$

i.e. it transforms like a field in the adjoint representation, with no inhomogeneous term.

The gauge-invariant action is formally almost identical to that of the $U(1)$ theory:

$$\mathcal{L} = i\bar{\psi}\, \not{D}\psi - m\bar{\psi}\psi - \frac{1}{2g^2}\mathrm{Tr}\, F_{\mu\nu}^2. \tag{2.16}$$

Here we have changed the letter we use to denote the coupling constant; we will usually reserve e for the electron charge, using g for a generic gauge coupling. Note also that it is necessary to take a trace of F^2 to obtain a gauge-invariant expression.

The matrix form for the fields may be unfamiliar, but it is very powerful. One can recover expressions in terms of more conventional fields by defining

$$A_\mu = A_\mu^a T_a, \tag{2.17}$$

where T_a are the group generators in representation appropriate to ψ. Then for $SU(N)$, for example, if the T_as are in the fundamental representation,

$$\mathrm{Tr}(T_a T_b) = \frac{1}{2}\delta_{ab}; \quad [T^a, T^b] = i f^{abc} T^c, \tag{2.18}$$

where f^{abc} are the structure constants of the group, and

$$A_\mu^a = 2\mathrm{Tr}(T_a A^\mu); \quad F_{\mu\nu}^a = \partial_\mu A_\nu^a - \partial_\nu A_\mu^a + f_{abc} A_\mu^a A_\nu^b. \tag{2.19}$$

While formally almost identical, there are great differences between the Abelian and non-Abelian theories. Perhaps the most striking is that the equations for the A_μs are non-linear in non-Abelian theories. This behavior means that, unlike the case of Abelian gauge fields, a theory of non-Abelian fields without matter is a non-trivial, interacting theory with interesting properties. With and without matter fields, this will lead to much richer behavior even classically. For example, we will see that non-Abelian theories sometimes contain solitons, localized finite-energy solutions of the classical equations. The most interesting of these are magnetic monopoles. At the quantum level, these non-linearities lead to properties such as asymptotic freedom and confinement.

As we have written the action, the matter fields ψ can appear in any representation of the group; one just needs to choose the appropriate matrices T^a. We can also consider scalars, as well as fermions. For a scalar field, ϕ, we define the covariant derivative, $D_\mu\phi$ as before, and add to the action a term $|D_\mu\phi|^2$ for a complex field, or $(1/2)(D_\mu\phi)^2$ for a real field.

2.2 Realizations of symmetry in quantum field theory

The most primitive exercise we can do with the Yang–Mills Lagrangian is set $g = 0$ and examine the equations of motion for the fields A^μ. If we choose the gauge $\partial_\mu A^{\mu a} = 0$, all of the gauge fields obey

$$\partial^2 A^a_\mu = 0. \tag{2.20}$$

So, like the photon, all of the gauge fields, A^a_μ, of the Yang–Mills theory are massless. At first sight, there is no obvious place for these fields in either the strong or the weak interactions. But it turns out that in non-Abelian theories, the possible ways in which the symmetry may be realized are quite rich. The symmetry can be realized in terms of massless gauge bosons; this is known as the *Coulomb phase*. This possibility is not relevant to the Standard Model, but will appear in some of our more theoretical considerations later. A second is known as the *Higgs phase*. In this phase, the gauge bosons are massive. In the third, the *confinement phase*, there are no physical states with the quantum numbers of isolated quarks (particles in the fundamental representation), and the gauge bosons are also massive. The second phase is relevant to the weak interactions; the third, confinement phase, to the strong interactions.[1]

2.2.1 The Goldstone phenomenon

Before introducing the Higgs phase, it is useful to discuss global symmetries. While we will frequently argue, like Yang and Mills, that global symmetries are less fundamental than local ones, they are important in nature. Examples are isospin, the chiral symmetries of the strong interactions, and baryon number. We can represent the action of such a symmetry much as we represented the symmetry action in Yang–Mills theory:

$$\Phi \to g_\alpha \Phi \tag{2.21}$$

[1] The differences between the confinement and Higgs phases are subtle, as first stressed by Fradkin, Shenker and 't Hooft. But we now know that the Standard Model is well described by a weakly coupled field theory in the Higgs phase.

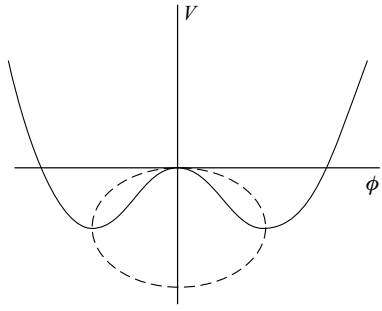

Fig. 2.1. Scalar potential with negative mass-squared. The stable minimum leads to broken symmetry.

where Φ is some set of fields and g is now a constant matrix, independent of space. Such symmetries are typically accidents of the low-energy theory. Isospin, for example, arises as we will see because the masses of the u and d quarks are small compared with other hadronic scales of quantum chromodynamics. Then g is the matrix:

$$g_{\vec{\alpha}} = e^{i\frac{\vec{\alpha}\cdot\vec{\sigma}}{2}} \tag{2.22}$$

acting on the u and d quark doublet. Note that $\vec{\alpha}$ is not a function of space, but it is a continuous parameter, so we will refer to such symmetries as continuous global symmetries. In the case of isospin, it is also important that the electromagnetic and weak interactions, which violate this symmetry, are small perturbations on the strong interactions.

The simplest model of a continuous global symmetry is provided by a complex field transforming under a $U(1)$ symmetry:

$$\phi \to e^{i\alpha}\phi. \tag{2.23}$$

We can take for the Lagrangian for this system:

$$\mathcal{L} = |\partial_\mu \phi|^2 - m^2|\phi|^2 - \frac{1}{2}\lambda|\phi|^4. \tag{2.24}$$

If $m^2 > 0$, and λ is small, this is just a theory of a weakly interacting, complex scalar. The states of the theory can be organized as states of definite $U(1)$ charge. This is the unbroken phase. On the other hand, m^2 is just a parameter, and we can ask what happens if $m^2 = -\mu^2 < 0$. In this case, the potential,

$$V(\phi) = -\mu^2|\phi|^2 + \lambda|\phi|^4, \tag{2.25}$$

looks as in Fig. 2.1. There is a set of degenerate minima,

$$\langle\phi\rangle_\alpha = \frac{\mu}{\sqrt{2\lambda}}e^{i\alpha}. \tag{2.26}$$

These classical ground states are obtained from one another by symmetry transformations. Classically the theory has a set of degenerate ground states, labeled by α; in somewhat more mathematical language, there is a manifold of vacuum states. Quantum mechanically, it is necessary to choose a particular value of α. As explained in the next section, if one chooses α, no local operator, e.g. no small perturbation, will take the system into another state of different α. To simplify the writing, take $\alpha = 0$. Then we can parameterize:

$$\phi = \frac{1}{\sqrt{2}}(v + \sigma(x))e^{i\pi(x)/v} \approx \frac{1}{\sqrt{2}}(v + \sigma(x) + i\pi(x)). \tag{2.27}$$

Here $v = \mu/\sqrt{\lambda}$ is known as the "vacuum expectation value" of ϕ. In terms of σ and π, the Lagrangian takes the form:

$$\mathcal{L} = \frac{1}{2}[(\partial_\mu\sigma)^2 + (\partial_\mu\pi)^2 - 2\mu^2\sigma^2 + \mathcal{O}(\sigma, \pi)^3]. \tag{2.28}$$

So we see that σ is an ordinary, real scalar field of mass-squared $2\mu^2$, while the π field is massless. The fact that the π is massless is not a surprise: the mass represents the energy cost of turning on a zero-momentum excitation of π, but such an excitation is just a symmetry transformation of ϕ, $v \to ve^{i\pi(0)}$. So there *is* no energy cost.

The appearance of massless particles when a symmetry is broken is known as the Nambu–Goldstone phenomenon, and π is called a Nambu–Goldstone boson. In any theory with scalars, a choice of minimum may break some symmetry. This means that there is a manifold of vacuum states. The broken symmetry generators are those which transform the system from one point on this manifold to another. Because there is no energy cost associated with such a transformation, there is a massless particle associated with each broken symmetry generator. This result is very general. Symmetries can be broken not only by expectation values of scalar fields, but by expectation values of composite operators, and the theorem holds. A proof of this result is provided in Appendix B. In nature, there are a number of excitations which can be identified as Goldstone or almost Goldstone ("pseudo-Goldstone") bosons. These include spin waves in solids and the pi mesons. We will have much more to say about the pions later.

2.2.2 Aside: choosing a vacuum

In quantum mechanics, there is no notion of a spontaneously broken symmetry. If one has a set of degenerate classical configurations, the ground state will invariably involve a superposition of these configurations. If we took σ and π to be functions only of t, then the σ–π system would just be an ordinary quantum-mechanical

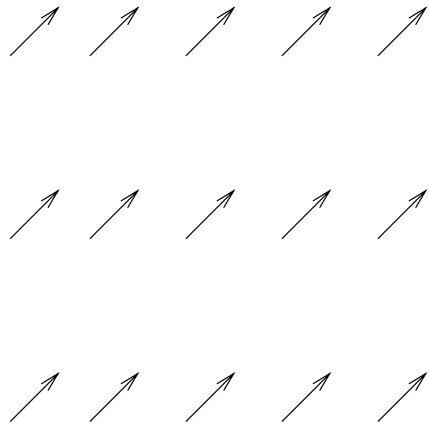

Fig. 2.2. In a ferromagnet, the spins are aligned, but the direction is arbitrary.

system with two degrees of freedom. Here σ would correspond to an anharmonic oscillator of frequency $\omega = \sqrt{2}\mu$. Placing this particle in its ground state, one would be left with the coordinate π. Note that π, in Eq. (2.27), is an angle, like the azimuthal angle, ϕ, in ordinary quantum mechanics. We could call its conjugate variable L_z. The lowest-lying state would be the zero-angular-momentum state, a uniform superposition of all values of π. In field theory at finite volume, the situation is similar. The zero-momentum mode of π is again an angular variable, and the ground state is invariant under the symmetry. But at infinite volume, the situation is different. One is forced to choose a value of π.

This issue is most easily understood by considering a different problem: rotational invariance in a magnet. Consider Fig. 2.2, where we have sketched a ferromagnet with spins aligned at an angle θ. We can ask: what is the overlap of two states, one with $\theta = 0$, one at θ, i.e. what is $\langle\theta|0\rangle$? For a single site, the overlap between the state $|+\rangle$ and the rotated state is:

$$\left\langle + \left| e^{i\tau_1\theta/2} \right| + \right\rangle = \cos(\theta/2). \tag{2.29}$$

If there are N such sites, the overlap behaves as

$$\langle\theta|0\rangle \sim (\cos(\theta/2))^N \tag{2.30}$$

i.e. it vanishes exponentially rapidly with the "volume," N.

For a continuum field theory, states with differing values of the order parameter, v, also have no overlap in the infinite volume limit. This is illustrated by the theory of a scalar field with Lagrangian:

$$\mathcal{L} = \frac{1}{2}(\partial_\mu\phi)^2. \tag{2.31}$$

For this system, there is no potential, so the expectation value, $\phi = v$, is not fixed. The Lagrangian has a symmetry, $\phi \to \phi + \delta$, for which the charge is just

$$Q = \int d^3x \, \Pi(\vec{x}) \tag{2.32}$$

where Π is the canonical momentum. So we want to study

$$\langle v|0\rangle = \langle 0|e^{iQ}|0\rangle. \tag{2.33}$$

We must be careful how we take the infinite-volume limit. We will insist that this be done in a smooth fashion, so we will define:

$$\begin{aligned} Q &= \int d^3x \, \partial_0 \phi e^{-\vec{x}^2/V^{2/3}} \\ &= -i \int \frac{d^3k}{(2\pi)^3} \sqrt{\frac{\omega_k}{2}} \left(\frac{V^{1/3}}{\sqrt{\pi}}\right)^3 e^{-\vec{k}^2 V^{2/3}/4} [a(\vec{k}) - a^\dagger(\vec{k})]. \end{aligned} \tag{2.34}$$

Now, one can evaluate the matrix element, using

$$e^{A+B} = e^A e^B \bar{e}^{(1/2)[A,B]}$$

(provided that the commutator is a c-number), giving

$$\langle 0|e^{iQ}|0\rangle = e^{-cv^2 V^{2/3}}, \tag{2.35}$$

where c is a numerical constant. So the overlap vanishes with the volume. You can convince yourself that the same holds for matrix elements of local operators. This result does not hold in $0+1$ and $1+1$ dimensions, because of the severe infrared behavior of theories in low dimensions. This is known to particle physicists as Coleman's theorem, and to condensed matter theorists as the Mermin–Wagner theorem. This theorem will make an intriguing appearance in string theory, where it is the origin of energy–momentum conservation.

2.2.3 The Higgs mechanism

Suppose that the $U(1)$ symmetry of the previous section is local. In this case, even a spatially varying $\pi(x)$ represents a symmetry transformation, and by a gauge choice it can be eliminated. In other words, by a gauge transformation, we can bring ϕ to the form:

$$\phi = \frac{1}{\sqrt{2}}(v + \sigma(x)). \tag{2.36}$$

In this gauge, the kinetic terms for ϕ take the form:

$$|D_\mu \phi|^2 = \frac{1}{2}(\partial_\mu \sigma)^2 + \frac{1}{2} A_\mu^2 v^2. \tag{2.37}$$

The last term is a mass term for the gauge field. To determine the actual value of the mass, we need to examine the kinetic term for the gauge fields:

$$-\frac{1}{2g^2}(\partial_\mu A^\nu)^2 + \cdots . \tag{2.38}$$

So the gauge field has a mass $m_A^2 = g^2 v^2$.

This phenomenon, that the gauge boson becomes massive when the gauge symmetry is spontaneously broken, is known as the Higgs mechanism. While formally quite similar to the Goldstone phenomenon, it is also quite different. The fact that there is no massless particle associated with motion along the manifold of ground states is not surprising – these states are all physically equivalent. Symmetry breaking, in fact, is a puzzling notion in gauge theories, since gauge transformations describe entirely equivalent physics (gauge symmetry is often referred to as a redundancy in the description of a system). Perhaps the most important lesson here is that gauge invariance does not necessarily mean, as it does in electrodynamics, that the gauge bosons are massless.

2.2.4 Goldstone and Higgs phenomena for non-Abelian symmetries

Both the Goldstone and Higgs phenomena generalize to non-Abelian symmetries. In the case of global symmetries, for every generator of a broken global symmetry, there is a massless particle. For local symmetries, each broken generator gives rise to a massive gauge boson.

As an example, relevant both to the strong and the weak interactions, consider a theory with a symmetry $SU(2)_{\rm L} \times SU(2)_{\rm R}$. Take M to be a Hermitian, matrix field,

$$M = \sigma + i\vec{\pi} \cdot \vec{\sigma}. \tag{2.39}$$

Under the symmetry, which we first take to be global, M transforms as

$$M \to g_{\rm L} M g_{\rm R} \tag{2.40}$$

with $g_{\rm L}$ and $g_{\rm R}$ $SU(2)$ matrices. We can take the Lagrangian to be

$$\mathcal{L} = \ {\rm Tr}\,(\partial_\mu M^\dagger \partial^\mu M) - V(M^\dagger M). \tag{2.41}$$

This Lagrangian respects the symmetry. If the curvature of the potential at the origin is negative, M will acquire an expectation value. If we take:

$$\langle M \rangle = \langle \sigma \rangle \tag{2.42}$$

then some of the symmetry is broken. However, the expectation value of M is invariant under the subgroup of the full symmetry group with $g_{\rm L} = g_{\rm R}$. In other words,

the unbroken symmetry is $SU(2)$. Under this symmetry, the fields $\vec{\pi}$ transform as a vector. In the case of the strong interactions, this unbroken symmetry can be identified with isospin. In the case of the weak interactions, there is an approximate global symmetry reflected in the masses of the W and Z particles, as we will discuss later.

2.2.5 Confinement

There is still another possible realization of gauge symmetry: confinement. This is crucial to our understanding of the strong interactions. As we will see, Yang–Mills theories, without too much matter, become weak at short distances. They become strong at large distances. This is just what is required to understand the qualitative features of the strong interactions: free quark and gluon behavior at very large momentum transfers, but strong forces at larger distances, so that there are no free quarks or gluons. As is the case in the Higgs mechanism, there are no massless particles in the spectrum of hadrons: QCD is said to have a "mass gap." These features of strong interactions are supported by extensive numerical calculations, but they are hard to understand through simple analytic or qualitative arguments (indeed, if you can offer such an argument, you can win one of the Clay prizes of $1 million). We will have more to say about the phenomenon of confinement when we discuss lattice gauge theories.

One might wonder: what is the difference between the Higgs mechanism and confinement? This question was first raised by Fradkin and Shenker and by 't Hooft, who also gave an answer: there is often no qualitative difference. The qualitative features of a Higgs theory like the weak interactions can be reproduced by a confined, strongly interacting theory. However, the detailed predictions of the weakly interacting Weinberg–Salaam theory are in close agreement with experiment, and those of the strongly interacting theory are not.

2.3 The quantization of Yang–Mills theories

In this book, we will encounter a number of interesting classical phenomena in Yang–Mills theory but, in most of the situations in nature which concern us, we are interested in the quantum behavior of the weak and strong interactions. Abelian theories such as QED are already challenging. One can perform canonical quantization in a gauge, such as Coulomb gauge or a light cone gauge, in which unitarity is manifest – all of the states have positive norm. But in such a gauge, the covariance of the theory is hard to see. Or one can choose a gauge where Lorentz invariance is manifest, but not unitarity. In QED it is not too difficult to show at the level of Feynman diagrams that these gauge choices are equivalent. In non-Abelian theories, canonical quantization is still more challenging. Path integral methods provide a much more powerful approach to the quantization of these theories.

A brief review of path integration appears in Appendix C. Here we discuss gauge fixing and derive the Feynman rules. We start with the gauge fields alone; adding the matter fields – scalars or fermions – is not difficult. The basic path integral is:

$$\int [dA_\mu] e^{iS}. \tag{2.43}$$

The problem is that this integral includes a huge redundancy: the gauge transformations. To deal with this, we want to make a gauge choice, for example:

$$G^a\big(A^a_\mu\big) = \partial_\mu A^{\mu a} = 0. \tag{2.44}$$

We can't simply insert this in the path integral without altering its value. Instead, we insert 1 in the form:

$$1 = \int [dg]\delta\big(G\big(A^g_\mu\big)\big)\Delta[A]. \tag{2.45}$$

Here we have reverted to our matrix notation: G is a general gauge fixing condition; A^g_μ denotes the gauge transform of A_μ by g. The Δ is a functional determinant, known as the Faddeev–Popov determinant. Note that Δ is gauge invariant, $\Delta[A^h] = \Delta[A]$. This follows from the definition:

$$\int [dg]\delta\big(G\big(A^{hg'}_\mu\big)\big) = \int [dg]\delta\big(G\big(A^{g'}_\mu\big)\big) \tag{2.46}$$

where, in the last step, we have made the change of variables $g \to h^{-1}g$. We can write a more explicit expression for Δ as a determinant. To do this, we first need an expression for the variation of the A under an infinitesimal gauge transformation. Writing $g = 1 + i\omega$, and using the matrix form for the gauge field,

$$\delta A_\mu = \partial_\mu \omega + i[\omega, A_\mu]. \tag{2.47}$$

This can be written elegantly as a covariant derivative of ω, where ω is thought of as a field in the adjoint representation:

$$\delta A_\mu = D_\mu \omega. \tag{2.48}$$

If we make the specific choice $G = \partial_\mu A^\mu$, to evaluate Δ we need to expand G about the field A_μ for which $G = 0$:

$$G(A + \delta A) = \partial_\mu D^\mu \omega = \partial^2 \omega + i[A_\mu, \partial_\mu \omega] \tag{2.49}$$

or, in the index form:

$$G\big(A^a_\mu\big) = (\partial^2 \delta^{ac} + f^{abc} A^{\mu\ b} \partial_\mu)\omega^c. \tag{2.50}$$

So

$$\Delta[A] = \det[\partial^2 \delta^{ac} + f^{abc} A^{\mu\ b} \partial_\mu]^{-1/2}. \tag{2.51}$$

We will discuss strategies to evaluate this determinant shortly.

At this stage, we have reduced the path integral to:

$$Z = \int [dA_\mu]\delta(G(A))\Delta[A]e^{iS} \tag{2.52}$$

and we can write down Feynman rules. The δ-function remains rather awkward to deal with, though, and this expression can be simplified through the following trick. Introduce a function, ω, and average over ω with a Gaussian weight factor:

$$Z = \int [d\omega]e^{i\int d^4x(\omega^2/\xi)} \sum \int [dA_\mu]\delta(G(A) - \omega)\Delta[A]e^{iS}. \tag{2.53}$$

We can do the integral over the δ-function. The quadratic terms in the exponent are now:

$$\int d^4x\ A^{\mu a}\left[-\partial^2\eta_{\mu\nu} + \partial_\mu\partial_\nu\left(1 - \frac{1}{\xi}\right)\right]A^{\nu b}. \tag{2.54}$$

We can invert to find the propagator. In momentum space:

$$D_{\mu\nu} = -\frac{\eta_{\mu\nu} + (\xi - 1)k_\mu k_\nu/k^2}{k^2 + i\epsilon}. \tag{2.55}$$

To write explicit Feynman rules, we need also to deal with the Faddeev–Popov determinant. Feynman long ago guessed that the unitarity problems of Yang–Mills theories could be dealt with by introducing fictitious scalar fields with the wrong statistics. Our expression for Δ can be reproduced by a functional integral for such particles:

$$\Delta = \int [dc^a][dc^{a\dagger}]\exp\left(i\int d^4x(c^{a\dagger}(\partial^2\delta^{ab} + f^{abc}A^{\mu\ c}\partial_\mu)c^b)\right). \tag{2.56}$$

From here, we can read off the Feynman rules for Yang–Mills theories, including matter fields. These are summarized in Fig. 2.3.

2.3.1 Gauge fixing in theories with broken gauge symmetry

Gauge fixing in theories with broken gauge symmetries raise some new issues. We consider first a $U(1)$ gauge theory with a single charged scalar field, ϕ. We suppose that the potential is such that $\langle\phi\rangle = v/\sqrt{2}$. We call e the gauge coupling and take the conventional scaling for the gauge kinetic terms. We can, again, parameterize the field as:

$$\phi = \frac{1}{\sqrt{2}}(v + \sigma(x))e^{(i\pi)/v}. \tag{2.57}$$

Then we can again choose a gauge in which $\pi(x) = 0$. This gauge is known as unitary gauge, since, as we have seen, in this gauge we have exactly the degrees of

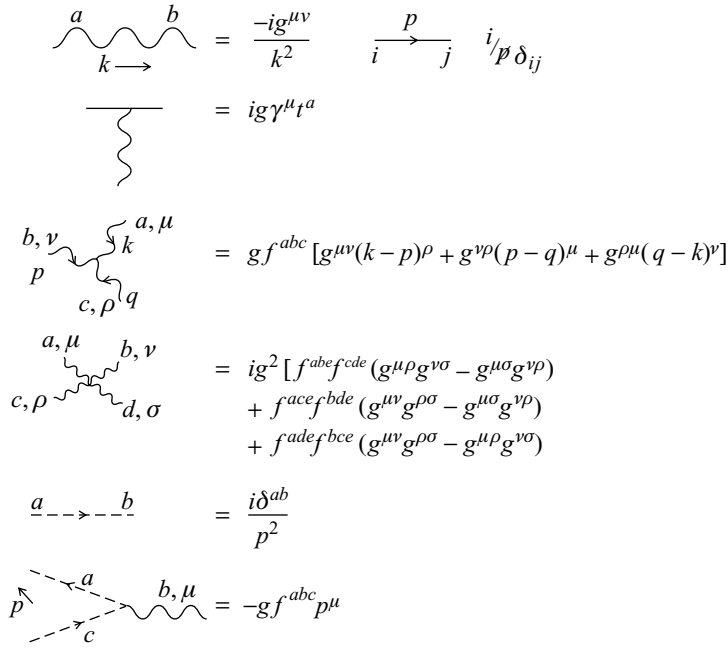

Fig. 2.3. Feynman rules for Yang–Mills theory.

freedom we expect physically: a massive gauge boson and a single real scalar. But this gauge is not convenient for calculations. The gauge boson propagator in this gauge is:

$$\langle A_\mu A_\nu \rangle = -\frac{i}{k^2 - M_V^2}\left(\eta_{\mu\nu} - \frac{k_\mu k_\nu}{M_V^2}\right). \tag{2.58}$$

Because of the factors of momentum in the second term, individual Feynman diagrams have bad high-energy behavior. A more convenient set of gauges, known as R_ξ gauges, avoids this difficulty, at the price of keeping the π field (sometimes misleadingly called the Goldstone particle) in the Feynman rules. We take, in the path integral, the gauge-fixing function:

$$G = \frac{1}{\sqrt{\xi}}(\partial_\mu A^\mu \xi - ev\pi(x)). \tag{2.59}$$

The extra term has been judiciously chosen so that when we exponentiate, the $A^\mu \partial_\mu \pi$ terms in the action cancel. Explicitly, we have:

$$\begin{aligned}\mathcal{L} = &-\frac{1}{2}A_\mu\left(\eta^{\mu\nu}\partial^2 - \left(1 - \frac{1}{\xi}\right)\partial^\mu\partial^\nu - (e^2v^2)\eta^{\mu\nu}\right)A_\nu \\ &+ \frac{1}{2}(\partial_\mu\sigma)^2 - \frac{1}{2}m_\sigma^2\sigma^2 + \frac{1}{2}(\partial_\mu\pi)^2 - \frac{\xi}{2}(ev)^2\pi^2 + \mathcal{O}(\phi^3).\end{aligned} \tag{2.60}$$

If we choose $\xi = 1$ ('t Hooft–Feynman gauge), the propagator for the gauge boson is simply

$$\langle A_\mu A_\nu \rangle = \frac{-i}{k^2 - M_V^2} \eta_{\mu\nu} \tag{2.61}$$

with $M^2 = e^2 v^2$, but we have also the field π explicitly in the Lagrangian, and it has propagator:

$$\langle \pi \pi \rangle = \frac{i}{k^2 - M_V^2}. \tag{2.62}$$

The mass here is just the mass of the vector boson (for other choices of ξ, this is not true).

This gauge choice is readily extended to non-Abelian theories, with similar results: the gauge bosons have simple propagators, like those of massive scalars multiplied by $\eta_{\mu\nu}$. The Goldstone bosons appear explicitly in perturbation theory, with propagators appropriate to massive fields. The Faddeev–Popov ghosts have couplings to the scalar fields.

2.4 The particles and fields of the Standard Model

We are now in a position to write down the Standard Model. It is amazing that, at a microscopic level, almost everything we know about nature is described by such a simple structure. The gauge group is $SU(3)_c \times SU(2)_L \times U(1)_Y$. The subscript c denotes color, L means "left-handed" and Y is called hypercharge. Corresponding to these different gauge groups, there are gauge bosons, A_μ^a, $a = 1, \ldots, 8$, W_μ^i, $i = 1, 2, 3$, and B_μ.

One of the most striking features of the weak interactions is the violation of parity. In terms of four-component fields, this means that factors of $(1 - \gamma_5)$ appear in the couplings of fermions to the gauge bosons. In such a situation, it is more natural to work with two-component spinors. For the reader unfamiliar with such spinors, a simple introduction appears in Appendix A. Such spinors are the basic building blocks of the four-dimensional spinor representations of the Lorentz group. All spinors can be described as two-component spinors, with various quantum numbers. For example, quantum electrodynamics, which is parity invariant with a massive fermion, can be described in terms of two left-handed fermions, e and $\bar{e}$, with electric charge $-e$ and $+e$, respectively. The Lagrangian takes the form:

$$\mathcal{L} = ie\sigma^\mu D_\mu e^* + i\bar{e}\sigma^\mu D_\mu \bar{e}^* - m\bar{e}e - m\bar{e}^* e^*. \tag{2.63}$$

The covariant derivatives are those appropriate to fields of charge e and $-e$. Parity is the symmetry $\vec{x} \to -\vec{x}$, $e \leftrightarrow \bar{e}^*$, and $\vec{A} \to -\vec{A}$.

We can specify the fermion content of the Standard Model by giving the gauge quantum numbers of the left-handed spinors. So, for example, there are quark doublets which are 3s of color and doublets of $SU(2)$, with hypercharge $1/3$: $Q = (3, 2)_{1/3}$. The appropriate covariant derivative is:

$$D_\mu Q = \left(\partial_\mu - i g_s A^a_\mu T^a - i g W^i_\mu T^i - i \frac{g'}{2}\frac{1}{3} B_\mu\right) Q. \tag{2.64}$$

Here T^i are the generators of $SU(2)$, $T^i = \sigma^i/2$. These are normalized as

$$\mathrm{Tr}(T^i T^j) = \frac{1}{2}\delta^{ij}. \tag{2.65}$$

The T^a are the generators of $SU(3)$; in terms of Gell-Mann's $SU(3)$ matrices, $T^a = \lambda^a/2$. They are normalized like the $SU(2)$ matrices, $\mathrm{Tr}\, T^a T^b = (1/2)\delta^{ab}$.

We have followed the customary definition, in coupling B_μ to half the hypercharge current. We have also scaled the fields so that the couplings appear in the covariant derivative, and labeled the $SU(3)_c$, $SU(2)_L$, and $U(1)_Y$ coupling constants as g_s, g and g', respectively. Using matrix-valued fields, defined with the couplings in front of the gauge kinetic terms, this covariant derivative can be written in a very compact manner:

$$D_\mu Q = \left(\partial_\mu - i A_\mu - i W_\mu - \frac{i}{2} \times \frac{1}{3} B_\mu\right) Q. \tag{2.66}$$

As another example, the Standard Model contains lepton fields, L, with no $SU(3)$ quantum numbers, but which are $SU(2)$ doublets with hypercharge -1. The covariant derivative is:

$$D_\mu L = \left(\partial_\mu - i g W^i_\mu T^i - \frac{g'}{2} B_\mu\right) L. \tag{2.67}$$

Similarly for the Higgs doublet, ϕ.

So we can summarize the fermion contact in the Standard Model with the following data:

	$SU(3)$	$SU(2)$	$U(1)_Y$
Q_f	3	2	1/3
$\bar{u}_f$	$\bar{3}$	1	$-4/3$
$\bar{d}_f$	$\bar{3}$	1	2/3
L_f	1	2	-1
$\bar{e}_f$	1	1	2

(2.68)

Here f labels the quark or lepton flavor, or generation number, $f = 1, 2, 3$. For example,

$$L_1 = \begin{pmatrix} \nu_e \\ e \end{pmatrix} \quad L_2 = \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix} L_3 = \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}. \tag{2.69}$$

Why there is this repetitive structure, these three generations, is one of the great puzzles of the Standard Model, to which we will return. In terms of these two-component fields (indicated generically by ψ_i), the gauge-invariant kinetic terms have the form:

$$\mathcal{L}_{f,k} = -i \sum_i \psi_i D_\mu \sigma^\mu \psi_i^* \tag{2.70}$$

where the covariant derivatives are those appropriate to the representation of the gauge group.

Unlike QED (where, in two-component language, parity interchanges e and $\bar{e}^*$), the model does not have a parity symmetry. The fields Q and $\bar{u}$, $\bar{d}$ transform under different representations of the gauge group. There is simply no discrete symmetry one can find which is the analog of the parity symmetry of QED.

In order to account for the masses of the W and Z bosons and the quarks and leptons, the simplest version of the Standard Model includes a scalar, ϕ, which transforms as a $(1, 2)_1$ of the Standard Model gauge group. This Higgs field possesses both self-couplings and Yukawa couplings to the fermions. Its kinetic term is simply

$$\mathcal{L}_{\phi,k} = |D_\mu \phi|^2. \tag{2.71}$$

The Higgs potential is similar to that of our toy model:

$$V(\phi) = \mu^2 |\phi|^2 + \lambda |\phi|^4. \tag{2.72}$$

This is completely gauge invariant. But if μ^2 is negative, the gauge symmetry breaks as before. We will describe this breaking, and the mass matrix of the gauge bosons, shortly. At this point, we have written the most general renormalizable self-couplings of the scalar fields. Renormalizability and gauge invariance permit one other set of couplings in the Standard Model: Yukawa couplings of the scalars to the fermions. The most general such couplings are:

$$\mathcal{L}_{\text{Yuk}} = y^U_{f,f'} Q_f \bar{u}_{f'} \sigma_2 \phi^* + y^D_{f,f'} Q_f \bar{d}_{f'} \phi + y^L_{f,f'} L_f \bar{e}_{f'} \phi. \tag{2.73}$$

Here y^U, y^D and y^L are general matrices in the space of flavors.

We can simplify the Yukawa coupling matrices significantly by redefining fields. Any 3×3 matrix can be diagonalized by separate left and right $U(3)$ matrices. To

see this, suppose one has some matrix, M, not necessarily Hermitian. The matrices:

$$A = MM^\dagger \quad B = M^\dagger M \tag{2.74}$$

are Hermitian matrices. A can be diagonalized by a unitary transformation, $U_{\rm L}$, say, and B by a unitary transformation $U_{\rm R}$. Then the matrices:

$$U_{\rm L} M U_{\rm R}^\dagger \quad U_{\rm R} M^\dagger U_{\rm L}^\dagger \tag{2.75}$$

are diagonal. So by redefining fields, we can simplify the matrices y^U, y^D and y^L; it is customary to take y^U and y^L diagonal. There is a conventional form for y^D (the Cabibbo–Kobayashi–Maskawa matrix) which we will describe in Section 3.2.

To summarize, the entire Lagrangian of the Standard Model consists of the following.

(1) Gauge-invariant kinetic terms for the gauge fields,

$$\mathcal{L}_a = -\frac{1}{4g_{\rm s}^2} G_{\mu\nu}^2 - \frac{1}{4g^2} W_{\mu\nu}^2 - \frac{1}{4g'^2} F_{\mu\nu}^2 \tag{2.76}$$

(here we have returned to our scaling with the couplings in front; $G_{\mu\nu}$, $W_{\mu,\nu}$, and $F_{\mu\nu}$ are the $SU(3)$, $SU(2)$ and $U(1)$ field strengths).

(2) Gauge-invariant kinetic terms for the fermion and Higgs fields, $\mathcal{L}_{f,k}$, $\mathcal{L}_{\phi,k}$.

(3) Yukawa couplings of the fermions to the Higgs field, $\mathcal{L}_{\rm Yuk}$.

(4) The potential for the Higgs field, $V(\phi)$.

If we require renormalizability, i.e. we require that all of the terms in the Lagrangian be of dimension four or less, this is all we can write. It is extraordinary that this simple structure incorporates over a century of investigation of elementary particles.

2.5 The gauge boson masses

The field ϕ has an expectation value, which we can take to have the form:

$$\langle \phi \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} \tag{2.77}$$

where $v = \mu/\sqrt{\lambda}$. Expanding around this expectation value, the Higgs field can be written as:

$$\phi = e^{i\vec{\pi}(x)\cdot\vec{\sigma}/2v} \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + \sigma(x) \end{pmatrix}. \tag{2.78}$$

By a gauge transformation, we can set $\vec{\pi} = 0$. Not all of the gauge symmetry is broken by $\langle\phi\rangle$. Note that $\langle\phi\rangle$ is invariant under the $U(1)$ symmetry generated by

$$Q = T_3 + \frac{Y}{2}. \tag{2.79}$$

This is the electric charge. If we write:

$$L = \begin{pmatrix} \nu \\ e \end{pmatrix} \quad Q = \begin{pmatrix} u \\ d \end{pmatrix} \tag{2.80}$$

then the ν has charge 0, and e has charge -1; u has charge $2/3$, d charge $-1/3$. The charges of the singlets also work out correctly.

With this gauge choice, we can examine the scalar kinetic terms in order to determine the gauge boson masses. Keeping only terms quadratic in fluctuating fields (σ and the gauge fields), these now have the form:

$$|D_\mu\phi|^2 = \frac{1}{2}(\partial_\mu\sigma^2) + (0\ v)\left(igW_\mu^i\frac{\sigma^i}{2} + \frac{ig'}{2}B_\mu\right)\left(-igW^{\mu j}\frac{\sigma^j}{2} - \frac{ig'}{2}B^\mu\right)\begin{pmatrix}0\\ v\end{pmatrix}. \tag{2.81}$$

It is convenient to define the complex fields:

$$W_\mu^\pm = \frac{1}{\sqrt{2}}\left(W_\mu^1 \pm iW_\mu^2\right) \tag{2.82}$$

These are states of definite charge, since they carry zero hypercharge and $T_3 = \pm 1$. In terms of these fields, the gauge boson mass and kinetic terms take the form:

$$\begin{aligned}\partial_\mu W_\nu^+\partial^\mu W^{\nu-} + \frac{1}{2}\partial_\mu W_\nu^3\partial^\mu W^{\nu 3} + \frac{1}{2}\partial_\mu B_\nu\partial^\mu B^\nu \\ + g^2v^2W_\mu^+W^{\mu-} + \frac{1}{2}v^2\left(gW_\mu^3 - g'B_\mu\right)^2.\end{aligned} \tag{2.83}$$

Examining the terms involving the neutral fields, B_μ and W_μ^3, it is natural to redefine:

$$A_\mu = \cos(\theta_w)B_\mu + \sin(\theta_w)W_\mu^3; \quad Z_\mu = \sin(\theta_w)B_\mu + \cos(\theta_w)W_\mu^3 \tag{2.84}$$

where

$$\sin(\theta_w) = \frac{g'}{\sqrt{g^2 + g'^2}} \tag{2.85}$$

is known as the Weinberg angle. The field A_μ is massless, while the Ws and Zs have masses:

$$M_W^2 = g^2v^2 \quad M_Z^2 = (g^2 + g'^2)v^2 = M_W^2/\cos^2(\theta_w). \tag{2.86}$$

We can immediately see that A_μ couples to the current:

$$\begin{aligned} j^\mu_{\rm em} &= g' \cos(\theta_{\rm w}) \frac{1}{2} j^Y_\mu + g \sin(\theta_{\rm w}) j^3_\mu \\ &= e \left[\frac{1}{2} j^Y_\mu + j^3_\mu \right], \end{aligned} \tag{2.87}$$

where

$$e = \frac{g g'}{\sqrt{g^2 + g'^2}} \tag{2.88}$$

is the electric charge. So A_μ couples precisely as we expect the photon to couple. $W^{\mu\pm}$ couple to the charged currents of the four-fermi theory. The Z boson couples to:

$$j^Z_\mu = -g' \sin(\theta_{\rm w}) \frac{1}{2} j^Y_\mu + g \cos(\theta_{\rm w}) j^3_\mu. \tag{2.89}$$

2.6 Quark and lepton masses

Substituting the expectation value for the Higgs fields into the expression for the quark and lepton Yukawa couplings, Eq. (2.73) leads directly to masses for the quarks and leptons. The lepton masses and the masses for the up quarks are immediate:

$$m_{{\rm U}f} = y_{{\rm U}f} \frac{v}{\sqrt{2}} \quad m_{{\rm E}f} = y_{{\rm E}f} \frac{v}{\sqrt{2}}. \tag{2.90}$$

So, for example, the Yukawa coupling of the electron is $(m_e \sqrt{2})/v$.

But the masses for the d quarks are somewhat more complicated. Because y_D is not diagonal, we have a matrix, in flavor space, for the d quark masses:

$$(m_D)_{ff'} = (y_D)_{ff'} \frac{v}{\sqrt{2}}. \tag{2.91}$$

As we have seen, any matrix can be diagonalized by separate unitary transformations acting on the left and the right. So we can diagonalize this by separate rotations of the D quarks (within the quark doublets) and the $\bar{D}$ quarks. The rotation of the $\bar{D}$ quarks is just a simple redefinition of these fields. But the rotation of the D quarks is more significant, since it does not commute with $SU(2)_{\rm L}$. In other words, the quark masses are not diagonal in a basis in which the W boson couplings are diagonal. The basis in which the mass matrix is diagonal is known as the mass basis (the corresponding fields are often called "mass eigenstates").

The unitary matrix acting on the D quarks is known as the Cabibbo–Kobayashi–Maskawa, or CKM, matrix. In terms of this matrix, the coupling of the quarks to

the $W^{\pm}$ fields can be written:

$$W_{\mu}^{-} U_{f} \sigma^{\mu} D_{f'}^{*} V_{ff'} + W_{\mu}^{+} D_{f} \sigma^{\mu} U_{f'}^{*} V_{f'f}^{*}. \tag{2.92}$$

There are a variety of parameterizations of V, which we will discuss shortly. One interesting feature of the model is the Z couplings. Because V is unitary, these are diagonal in flavor. This explains why Z bosons don't mediate processes which change flavor, such as $K_{\mathrm{L}} \to \mu^{+}\mu^{-}$. The suppression of these *flavor-changing neutral currents* was one of the early, and critical, successes of the Standard Model.

Suggested reading

There are a number of textbooks with good discussions of the Standard Model, including those of Peskin and Schroeder (1995), Weinberg (1995), Cottingham and Greenwood (1998), Donoghue *et al.* (1992) and Seiden (2005). We cannot give a full bibliography of the Standard Model here, but the reader may want to examine some of the original papers, including: Yang and Mills (1954), Salam and Ward (1964), Weinberg (1967) and Glashow *et al.* (1970) on the weak interaction theory; 't Hooft (1971), Gross and Wilczek (1973) and Politzer (1973) on asymptotic freedom of the strong interactions. For discussion of the phases of gauge theories, see 't Hooft (1980) and Fradkin and Shenker (1979).

Exercises

(1) Georgi–Glashow Model: consider a gauge theory based on $SU(2)$, with Higgs field, $\vec{\phi}$, in the adjoint representation. Assuming that ϕ obtains an expectation value, determine the gauge boson masses. Identify the photon and the $W^{\pm}$ bosons. Is there a candidate for the Z boson?

(2) Consider the Standard Model with two generations. Show that there is no CP violation, and show that the KM matrices can be described in terms of a single angle, known as the Cabibbo angle.

3

Phenomenology of the Standard Model

The predictions of the Standard Model have been subjected to experimental tests in a broad range of processes. In processes involving leptons alone, and hadrons at high-momentum transfers, detailed, precise predictions are possible. In processes involving hadrons at low momentum, it is often possible to make progress using symmetry arguments. In still other cases, one can at least formulate a qualitative picture. In recent years, developments in lattice gauge theory have begun to offer the promise of reliable and precise predictions for at least some features of the large-distance behavior of hadrons. There exist excellent texts and reviews treating all of these topics. Here we will give only a brief survey, attempting to introduce ideas and techniques which are important in understanding what may lie beyond the Standard Model.

3.1 The weak interactions

We are now in a position to describe the weak interactions within the Standard Model. Summarizing our results for the W and Z masses, we have, at the tree level:

$$M_W^2 = \frac{\pi\alpha}{\sqrt{2}G_{\rm F}\sin^2\theta_{\rm w}} \quad M_Z^2 = \frac{\pi\alpha}{\sqrt{2}G_{\rm F}\sin^2\theta_{\rm w}\cos^2\theta_{\rm w}}, \tag{3.1}$$

where α is the fine-structure constant. Note, in particular, that in the leading approximation,

$$\frac{M_W^2}{M_Z^2} = \cos^2\theta_{\rm w}. \tag{3.2}$$

In these expressions, the Fermi constant is related to the W mass and the gauge coupling through:

$$G_{\rm F} = \sqrt{2}\frac{g^2}{8M_W^2} \quad G_{\rm F} = 1.166 \times 10^{-5}\ {\rm GeV}^{-2}. \tag{3.3}$$

The Weinberg angle is:

$$\sin^2(\theta_{\rm w}) = 0.231\,20(15). \tag{3.4}$$

The measured values of the W and Z masses are:

$$M_W = 80.425(38)\ \text{GeV} \quad M_Z = 91.1876(21)\ \text{GeV}. \tag{3.5}$$

One can see that the experimental quantities satisfy the theoretical relations to good accuracy. They are all in agreement at the part in 10^2–10^3 level when radiative corrections are included.

The effective Lagrangian for the quarks and leptons obtained by integrating out the W and Z particles is:

$$\mathcal{L}_W + \mathcal{L}_Z = \frac{8G_{\rm F}}{\sqrt{2}}\left[\left(J_\mu^1\right)^2 + \left(J_\mu^2\right)^2 + \left(J_\mu^3 - \sin^2\theta_{\rm w} J_{\mu\rm EM}\right)^2\right]. \tag{3.6}$$

The first two terms correspond to the exchange of the charged $W^\pm$ fields. The last term represents the effect of Z boson exchange. This structure has been tested extensively. In many experiments, the precision is sufficiently great that it is necessary to include radiative corrections.

The most precise tests of the weak interaction theory involve the Z bosons. Experiments at the LEP accelerator at CERN and the SLD accelerator at SLAC produced millions of Zs. These large samples permitted high-precision studies of the line shape and of the branching ratios to various final states. Care is needed in calculating radiative corrections. It is important to make consistent definitions of the various quantities. Detailed comparisons of theory and experiment can be found on the web site of the Particle Data Group (http://pdg.lbl.gov). As inputs, one generally takes the value of $G_{\rm F}$ measured in μ decays; the measured mass of the Z, and the fine structure constant. Outputs include the Z boson total width:

$$\textbf{experiment}\ \Gamma_Z = 2.4952 \pm 0.0023 \quad \textbf{theory}\ \Gamma_Z = 2.4956 \pm 0.0007. \tag{3.7}$$

The decay width of the Z to hadrons and leptons is also in close agreement (see Fig. 3.1). The W mass can also be computed with these inputs, and is by now measured quite precisely:

$$\textbf{experiment}\ M_W = 80.454 \pm 0.059 \pm 0.0023 \quad \textbf{theory}\ M_W = 80.390 \pm 0.18. \tag{3.8}$$

As of this writing, one piece of the Standard Model is missing: the Higgs boson. Taking seriously the simplest version of symmetry breaking in the Standard Model, there is one additional physical state. Its mass is a parameter (e.g. knowing the value of $G_{\rm F}$, it can be determined in terms of the Higgs quartic coupling, λ). Finding the Higgs particle is a serious experimental challenge, as its couplings to most particles

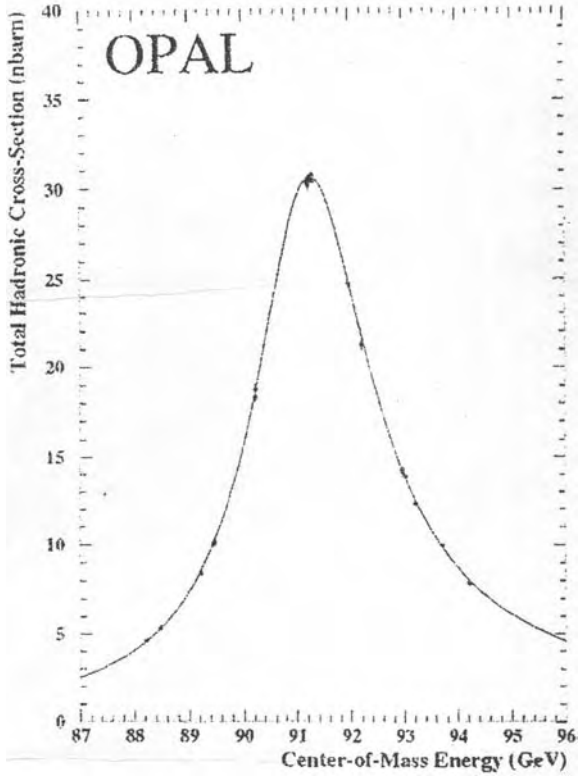

Fig. 3.1. OPAL results for the Z line shape. The solid line is theory; the dots are data (the size of the dots corresponds to the size of the error bars).

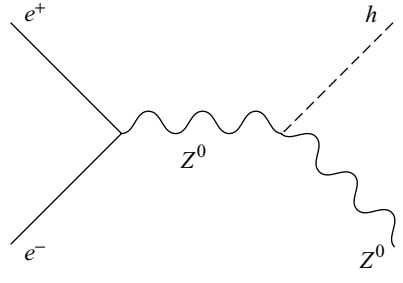

Fig. 3.2. Higgs can be produced in e^+e^- annihilation in association with a Z^0 particle.

are quite small. For example, couplings to leptons are suppressed by m_ℓ/v. The LEP experiments set a lower limit on the Higgs mass of about 115 GeV. The principal production mechanism is indicated in Fig. 3.2. For Higgs in the mass range explored at LEP, the most important decay channel involves decays to b quarks. At the LHC, other production mechanisms are dominant. For example, for a broad range of

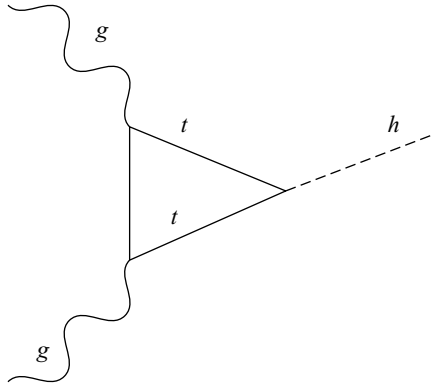

Fig. 3.3. In hadron colliders, Higgs particles can be produced by several mechanisms. The diagram above illustrates production by gluons.

masses, colliding gluons produce a virtual top pair, which in turn couple to the Higgs (Fig. 3.3).

3.2 The quark and lepton mass matrices

We have seen that we can take the Yukawa couplings for the up-type quarks to be diagonal, but cannot simultaneously diagonalize those for the down-type quarks. As a result, when the Higgs field gets an expectation value, the up quark masses are given by:

$$m_{uf} = \frac{(y_{\rm u})_f}{\sqrt{2}} v. \tag{3.9}$$

These are automatically diagonal. But the down quark masses are described by a 3×3 mass matrix,

$$m_{dff'} = \frac{(y_{\rm d})_{ff'}}{\sqrt{2}} v. \tag{3.10}$$

We can diagonalize this matrix by separate unitary transformations of the $\bar{d}$ and d fields. Because the d quarks are singlets of $SU(2)$, the transformation of $\bar{d}$ leaves the kinetic terms and gauge interactions of these quarks unchanged. But the transformation of the d quarks does not commute with $SU(2)$, so the couplings of the gauge bosons to these quarks are more complicated. The unitary transformation between the "mass" or "flavor" eigenstates and the weak interaction eigenstates is known as the CKM matrix. Denoting the mass eigenstates as u', d', etc., the

transformation has the form:

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix}. \tag{3.11}$$

There are various ways of parameterizing the CKM matrix. The Particle Data Group favors the following convention:

$$V = \begin{pmatrix} c_{12}c_{13} & s_{12}c_{13} & s_{13}e^{-i\delta_{13}} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}e^{i\delta_{13}} & c_{12}c_{23} - s_{12}c_{23}s_{13}e^{i\delta_{13}} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta_{13}} & -c_{12}s_{23} - s_{12}c_{23}s_{13}e^{i\delta_{13}} & c_{23}c_{13} \end{pmatrix}. \tag{3.12}$$

A number of features of this parameterization are worthy of note. First, V is unitary, $V^\dagger V = 1$. Second, V is real unless δ is non-zero. Thus δ provides a measure of CP violation. Finally, the parameterization has been chosen in such a way that, experimentally, all of the angles are small. As a result, all of the off-diagonal components of the matrix are small. The current measured values of the CKM angles are:

$$\begin{aligned} s_{12} &= 0.2243 \pm 0.0016 \quad & s_{23} &= 0.0413 \pm 0.0015 \\ s_{13} &= 0.0039 \pm 0.0005 \quad & \delta_{13} &= 1.05 \pm 0.024. \end{aligned} \tag{3.13}$$

Examining these quantities, one sees that not only are the off-diagonal matrix elements small, but they are hierarchically small. Wolfenstein has developed a convenient parameterization:

$$V = \begin{pmatrix} 1 - \lambda^2/2 & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda & 1 - \lambda^2/2 & A\lambda^2 \\ A\lambda^3(1 - \rho - i\eta) & -A\lambda^2 & 1 \end{pmatrix} + \mathcal{O}(\lambda^4). \tag{3.14}$$

Note that δ_{13}, and hence ρ and η, are not small. But there are several small entries, and a hierarchical structure reminiscent of the quark masses themselves.

From unitarity follow a number of relations among the elements of the matrix. For example,

$$V_{ud}V_{ub}^* + V_{cd}V_{cb}^* + V_{td}V_{tb}^* = 0. \tag{3.15}$$

From $V_{ud} \approx V_{cb} \approx V_{tb} \approx 1$, this becomes a relation between three complex numbers which says that they form a triangle – the *unitarity triangle*. Determining from experiment that these quantities do form a triangle is an important test of this model for the quark masses.

We should also give the values of the quark masses themselves. This is somewhat subtle, since in QCD it is necessary to specify an energy scale, much as one must specify the scale of the gauge coupling in QCD. With a scale of order 1 GeV, the

quark masses are roughly:

$$\begin{array}{lll} m_u \approx 1.5 - 4\ \text{MeV} & m_d \approx 4 - 8\ \text{MeV} & m_s \approx 80 - 130\ \text{MeV} \\ m_c \approx 1.15 - 1.35\ \text{GeV} & m_b \approx 4.1 - 4.4\ \text{GeV} & m_t \approx 174.3 \pm 5\ \text{GeV}. \end{array} \tag{3.16}$$

Before considering the small neutrino masses, the lepton Yukawa couplings can simply be taken diagonal, and there is no mixing. The lepton masses are:

$$m_e = 0.511\ \text{MeV} \quad m_\mu = 113\ \text{MeV} \quad m_\tau = 1.777\ \text{GeV}\,. \tag{3.17}$$

Overall, the picture of quark and lepton masses is quite puzzling. They vary over nearly five orders of magnitude. Correspondingly, the dimensionless Yukawa couplings have widely disparate values. Understanding this might well be a clue to what lies beyond the Standard Model.

3.3 The strong interactions

The strong interactions, as their name implies, are characterized by strong coupling. As a result, perturbative methods are not suitable for most questions. In comparing theory and experiment, it is necessary to focus on a few phenomena which are accessible to theoretical analysis. By itself, this is not particularly disturbing. A parallel with the quantum mechanics of electrons interacting with nuclei is perhaps helpful. We can understand simple atoms in detail; atoms with very large Z can be treated by Hartree–Fock or other methods. But atoms with intermediate Z can be dealt with, at best, by detailed numerical analysis accompanied by educated guesswork. Molecules are even more problematic, not to mention solids. But we are able to make detailed tests of the theory (and its extension in quantum electrodynamics) from the simpler systems, and develop qualitative understanding of the more complicated systems. In many cases, we can do quantitative analysis of the small fluctuations about the ground states of the complicated system.

In the theory of strong interactions, as we will see, many problems are hopelessly complicated. Low-lying spectra are hard; detailed exclusive cross sections in high-energy scattering essentially impossible. But there are many questions we can answer. Rates for many inclusive questions at very high energy and momentum transfer can be calculated with high precision. Qualitative features of the low lying spectrum of hadrons and their interactions at low energies can be understood in a qualitative (and sometimes quantitative) fashion by symmetry arguments. Recently, progress in lattice gauge theory has made it possible to perform calculations which previously seemed impossible, for features of spectra and even for interaction rates important for understanding the weak interactions.

3.3.1 Asymptotic freedom

The coupling of a gauge theory (and of a field theory more generally) is a function of energy or length scale. If a typical momentum transfer in a process is q, and if M denotes the cutoff scale,

$$\frac{8\pi^2}{g^2(q^2)} = \frac{8\pi^2}{g^2(\Lambda)} + b_0 \ln\left(\frac{q^2}{M^2}\right). \tag{3.18}$$

Here

$$b_0 = \frac{11}{3}C_{\rm A} - \frac{2}{3}c_i n_f^{(i)} - \frac{1}{3}c_i n_\phi^{(i)}. \tag{3.19}$$

In this expression, $n_f^{(i)}$ is the number of left-handed fermions in the ith representation, while $n_\phi^{(i)}$ is the number of scalars. $C_{\rm A}$ is the quadratic Casimir of the adjoint representation, and c_i the quadratic Casimir of the ith representation,

$$f^{acd}f^{bcd} = C_{\rm A}\delta^{ab} \quad {\rm Tr}(T^a T^b) = c_i\delta^{ab}. \tag{3.20}$$

These formulas are valid if the masses of the fermions and scalars are negligible at scale q^2. For example, in QCD, at scales of order the Z boson mass, the masses of all but the top quark can be neglected. All the quarks are in the fundamental representation, and there are no scalars. So $b_0 = 22/3$. As a result, g^2 gets smaller as q^2 gets larger, and, conversely, g^2 gets larger as q^2 gets smaller. Since momentum transfer is inversely proportional to a typical distance scale, one can say that the strong force gets weaker at short distances, and stronger at large distances. We will calculate b_0 in Section 3.5.

This is quite striking. In the case of QCD, it means that hadrons, when probed at very large momentum transfer, behave as collections of free quarks and gluons. Perturbation theory can be used to make precise predictions. On the other hand, viewed at large distances, hadrons are strongly interacting entities. Perturbation theory is not a useful tool, and other methods must be employed. The most striking phenomena in this regime are confinement – the fact that one cannot observe free quarks – and, closely related, the existence of a mass gap. Neither of these phenomena can be observed in perturbation theory.

3.4 The renormalization group

In thinking about physics beyond the Standard Model, by definition, we are considering phenomena involving degrees of freedom to which we have, as yet, no direct experimental access. The question of degrees of freedom which are as yet unknown is the heart of the problem of renormalization. In the early days of quantum field theory, it was often argued that one should be able to take a formal limit of infinite

cutoff, $\Lambda \to \infty$. Ken Wilson promulgated a more reasonable view: real quantum field theories describe physics below some characteristic scale, Λ. In a condensed matter system, this might be the scale of the underlying lattice, below which the system may often be described by a continuum quantum field theory. In the Standard Model, a natural scale is the scale of the W and Z bosons. Below this scale, the system can be described by a renormalizable field theory, QED plus QCD, along with certain non-renormalizable interactions – the four-fermi couplings of the weak interactions. In defining this theory, one can take the cutoff to be, say, M_W, or one can take it to be aM_W, for some $a < 1$. Depending on the choice of a, the values of the couplings will vary. The parameters of the low energy effective Lagrangian must depend on a in such a way that physical quantities are independent of this choice. The process of determining the values of couplings in an effective theory which reproduce the effects of some more microscopic theory is often referred to as *matching*.

Knowing how physical couplings depend on the cutoff, one can determine how physical quantities behave in the long-wavelength, infrared regime by simple dimensional analysis. Quantities associated with operators of dimension less than four will grow in the infrared. They are said to be "relevant." Those with dimension four will vary as powers of logarithms; they are said to be "marginal." Quantities with dimension greater than four, those conventionally referred to as "non-renormalizable operators," will be less and less important as the energy is lowered. They are said to be "irrelevant." In strongly interacting theories, the dimensions of operators can be significantly different than expected from naive classical considerations. The classification of operators as relevant, marginal, and irrelevant applies to their quantum behavior.

At sufficiently low energies, we can ignore the irrelevant, non-renormalizable couplings. Alternatively, by choosing the matching scale, M, low enough, only the marginal and relevant couplings are important. In a theory with only dimensionless couplings, the variation of the coupling with q^2 is closely related to its variation with the cutoff, M. Physical quantities are independent of the cutoff, so any explicit dependence on the cutoff must be compensated by the dependence of the couplings on M. On dimensional grounds, M must appear with q^2, so knowledge of dependence of couplings with M permits a derivation of their dependence on q^2. More precisely, in studying, say, a cross section, any explicit dependence on the cutoff must be compensated by the dependence of the coupling on the cutoff. Calling the physical quantity σ, we can express this as a differential equation, the *renormalization group equation*:

$$\left(M \frac{\partial}{\partial M} + \beta(g) \frac{\partial}{\partial g} \right) \sigma = 0. \tag{3.21}$$

Here the beta function (or "β-function") is given by:

$$\beta(g) = M\frac{\partial}{\partial M}g. \tag{3.22}$$

We can evaluate the beta function from our explicit expression, Eq. (3.18), for g^2:

$$\beta(g) = -b_0\frac{g^2}{16\pi^2}g. \tag{3.23}$$

We will compute b_0 in the next section. This equation has corrections in each order of perturbation theory and beyond.

So far we have expressed the coupling in terms of a cutoff and a physical scale. In an old-fashioned language, the coupling, $g^2(M)$, is the "bare" coupling. We can define a "renormalized coupling" at a scale μ^2, $g^2(\mu)$:

$$\frac{8\pi^2}{g^2(\mu^2)} = \frac{8\pi^2}{g^2(M)} + b_0 \ln\left(\frac{\mu}{M}\right). \tag{3.24}$$

In practice, it is necessary to give a more precise definition. We will discuss this when we compute the beta function in the next section. Because of this need to give a precise definition of the renormalized coupling, care is required in comparing theory and experiment. There are, as we will review shortly, a variety of definitions in common use, and it is important to be consistent.

Quantities like Green's functions are not physical, and obey an inhomogeneous equation. One can obtain this equation in a variety of ways. For simplicity, consider first a Green function with n scalar fields, such as:

$$G(x_1, \ldots, x_n) = \langle\phi(x_1)\ldots\phi(x_n)\rangle. \tag{3.25}$$

This Green function is related to the *renormalized* Green function as follows. If the theory is defined at a scale μ, the effective Lagrangian takes the form:

$$\mathcal{L}_\mu = Z^{-1}(\mu)(\partial_\mu\phi^2) + \cdots. \tag{3.26}$$

Here Z^{-1} arises from integrating out physics above the scale μ. It will typically include ultraviolet-divergent loop effects. Rescaling ϕ so that the kinetic term is canonical, $\phi = Z^{1/2}\phi_{\rm r}$, we have that

$$G(x_1, \ldots, x_n) = Z(\mu)^{n/2}G_{\rm r}(x_1, \ldots, x_n). \tag{3.27}$$

The left-hand side is independent of μ, so we can write an equation for $G_{\rm r}$:

$$\left(\mu\frac{\partial}{\partial\mu} + \beta(g)\frac{\partial}{\partial g} + n\gamma\right)G_{\rm r} = 0, \tag{3.28}$$

where γ, known as the *anomalous dimension*, is given by:

$$\gamma = \frac{1}{2}\mu\frac{\partial}{\partial\mu}\ln(Z). \tag{3.29}$$

In the case of several different fields, e.g. gauge fields, fermions and scalars, this equation is readily generalized. There is an anomalous dimension for each field, and the $n\gamma$ term is replaced by the appropriate number of fields of each type and their anomalous dimensions.

The effective action obeys a similar equation. Starting with:

$$\Gamma(x_1, \ldots, x_n) = Z(\mu)^{-n/2}\Gamma_{\rm r}(x_1, \ldots, x_n). \tag{3.30}$$

$$\left(\mu\frac{\partial}{\partial\mu} + \beta(g)\frac{\partial}{\partial g} - n\gamma\right)\Gamma_{\rm r} = 0. \tag{3.31}$$

These equations are readily solved. We could write the solution immediately, but an analogy with the motion of a fluid is helpful. A typical equation, for example, for the density of a component of the fluid (e.g. the density of bacteria) would take the form:

$$\left[\frac{\partial}{\partial t} + v(x)\frac{\partial}{\partial x} - \rho(x)\right]D(t, x) = 0, \tag{3.32}$$

where $D(t, x)$ is the density as a function of position and time, and $v(x)$ is the velocity of the fluid at x; ρ represents a source term (e.g. growth due to the presence of yeast, or a variable temperature). To solve this equation, one first solves for the motion of an element of fluid initially at x, i.e. one solves:

$$\frac{d}{dt}\bar{x}(t; x) = v(\bar{x}(t; x)) \quad \bar{x}(0; x) = x. \tag{3.33}$$

In terms of $\bar{x}$, we can immediately write down a solution for D:

$$\begin{aligned} D(t, x) &= D_0(\bar{x}(t; x))e^{\int_0^t dt'\rho(\bar{x}(t';x))} \\ &= D_0(\bar{x}(t; x))\, e^{\int_{\bar{x}(t)}^{x} dx' \frac{\rho(x')}{v(x')}}. \end{aligned} \tag{3.34}$$

Here D_0 is the initial density. One can check this solution by plugging in directly, but each piece has a clear physical interpretation. For example, if there were no source ($\rho = 0$), the solution just would become $D_0(\bar{x}(t; x))$. With no velocity, the source would lead to just the expected growth of the density.

Let's apply this to Green's functions. Consider, for example, a two-point function, $G(p) = ih(p^2/\mu^2)/p^2$. In our fluid dynamics analogy, the coupling, g, is the analog of the velocity; the log of the scale, $t = \ln(p/\mu)$, plays the role of the time. The equation for g is then:

$$\left[\frac{\partial}{\partial t} - \beta(g)\frac{\partial}{\partial g} - 2\gamma(g)\right]h(t) = 0. \tag{3.35}$$

Define $\bar{g}(\mu)$ as the solution of

$$\mu \frac{\partial}{\partial \mu} \bar{g}(\mu) = \beta(\bar{g}). \tag{3.36}$$

At lowest order, this is solved by Eq. (3.24). Then

$$h(p, g) = h(\bar{g}(t)) e^{2 \int_{t_0}^{t} dt' \gamma(\bar{g}(t', g))/\beta(\bar{g}(t', g))}. \tag{3.37}$$

One can write the solution in the form:

$$G(p, \lambda) = \frac{i}{p^2} G(\bar{g}(t; g)) e^{2 \int_{g}^{\bar{g}} dg' \frac{\gamma(g')}{\beta(g')}}. \tag{3.38}$$

3.5 Calculating the beta function

In the previous section, we presented the one loop result for the beta function and used it in various applications. In this section, we actually compute the beta function. There are a variety of ways to compute the variation of the gauge coupling with energy scale. One is to compute the potential for a very heavy quark–antiquark pair as a function of their separation (we use the term quark here loosely for a field in the N representation of $SU(N)$). The potential is a renormalization-group invariant quantity. At lowest order it is given by:

$$V(R) = -\frac{g^2 C_{\mathrm{F}}}{R} \tag{3.39}$$

where

$$C_{\mathrm{F}} = \sum_{a=1}^{N^2-1} T^a T^a. \tag{3.40}$$

The potential is a physical quantity; this is why it is renormalization-group invariant. In perturbation theory, it has corrections behaving as $g^2(\Lambda) \ln(R\Lambda)$. This follows simply from dimensional analysis. So if we choose $R = \Lambda$, the logarithmic terms disappear and we have:

$$V(R) = -g^2(R) \frac{C_F}{R} (1 + \mathcal{O}(g^2(R))). \tag{3.41}$$

In an asymptotically free theory like QCD, where the coupling gets smaller with distance, Eq. (3.41) becomes more and more reliable as R gets smaller. This result has physical applications. In the case of a bound state of a top quark and antiquark, one might hope that this would be a reasonable approximation, and describe the binding of the system. Taking $\alpha_s(R) \sim 0.1$, for example, would give a typical radius of order $(17\ \mathrm{GeV})^{-1}$, a length scale where one might expect perturbation theory to be reliable (and for which $\alpha_s(R) \sim 0.1$). By analogy with the hydrogen atom, one

would expect the binding energy to be of order 2 GeV. In practice, however, this is not directly relevant, since the width of the top quark is of the same order – the top quark decays before it has time to form a bound state. Still, it should be possible to see evidence for such QCD effects in production of $t\bar{t}$ pairs near threshold in e^+e^- annihilation.

A second approach is to study Green's functions in momentum space. The calculation is straightforward, if slightly more tedious than the analogous calculation in a $U(1)$ gauge theory (QED). The main complication is the three gauge boson vertex, which has many terms (at one loop, one can use symmetries to greatly simplify the algebra). It is necessary to have a suitable regulator for the integrals. By far the most efficient is the dimensional regularization technique of 't Hooft and Veltman. Here one initially allows the space-time dimensionality, d, to be arbitrary, and takes $d \to 4 - \epsilon$. For convenience, we include the two most frequently needed integration formulas below; their derivation can be found in many textbooks.

$$\int \frac{d^d k}{(k^2 + M^2)^n} = \frac{\pi^{d/2}\Gamma(n - d/2)}{\Gamma(n)} (M^2)^{d/2-n} \tag{3.42}$$

$$\int \frac{d^d k\, k^2}{(k^2 + M^2)^n} = \frac{\pi^{d/2}\Gamma(n - d/2 - 1)}{\Gamma(n)} (M^2)^{d/2-n+1}. \tag{3.43}$$

Ultraviolet divergences, such as would occur for $n = 2$ in the first integral, give rise to poles in the limit $\epsilon \to 0$. If we were simply to cut off the integral at $k^2 = \Lambda^2$, we would find:

$$\int \frac{d^4 k}{(2\pi)^4} \frac{1}{(k^2 + M^2)^2} \approx \frac{1}{16\pi^2} \ln(\Lambda/M). \tag{3.44}$$

In dimensional regularization this behaves as:

$$\int \frac{d^4 k}{(2\pi)^4} \frac{1}{(k^2 + M^2)^2} = \frac{1}{16\pi^2} \Gamma\left(\frac{\epsilon}{2}\right) \approx \frac{1}{8\pi^2 \epsilon}. \tag{3.45}$$

So ϵ should be thought of as $\ln(\Lambda^2)$. The computation of the Yang–Mills beta function by studying momentum-space Feynman diagrams can be found in many textbooks, and is outlined in the exercises.

Here we follow a different approach, known as the background field method. This technique is closely tied to the path integral, which will play an important role in this book. It is also closely tied to the Wilsonian view of renormalization. We break up A into a long-wavelength part, and a shorter-wavelength, fluctuating quantum part:

$$A^\mu = \mathcal{A}^\mu + a^\mu. \tag{3.46}$$

We can think of $\mathcal{A}^\mu$ as corresponding to modes of the field with momenta below the scale q, and a^μ as corresponding to higher momenta. We wish to compute an effective action for $\mathcal{A}^\mu$, integrating out the high-momentum modes:

$$\int [d\mathcal{A}] \int [da] e^{iS(\mathcal{A},a)} = \int [d\mathcal{A}] e^{iS_{\text{eff}}(\mathcal{A})}. \tag{3.47}$$

In calculating the effective action, we are treating $\mathcal{A}^\mu$ as a fixed, classical background. In this approach, one can work entirely in Euclidean space, which greatly simplifies the calculation.

Our first task is to write $e^{iS(\mathcal{A},a)}$. For this purpose, it is convenient to suppose that $\mathcal{A}$ satisfies its equation of motion. (Otherwise, it is necessary to introduce a source for a). A convenient choice of gauge is known as background field gauge:

$$D_\mu a^\mu = 0, \tag{3.48}$$

where D_μ is the covariant derivative defined with respect to the background field $\mathcal{A}$. At one loop, we only need to work out the action to second order in the fluctuating fields a^μ, ψ, ϕ. Consider, first, the fermion action. To quadratic order, we can set $a^\mu = 0$ in the Dirac Lagrangian. The same holds for scalars. So from the fermions and scalars we obtain:

$$\det(\not{D})^{n_f} \det(D^2)^{-n_\phi/2}. \tag{3.49}$$

The fermion functional determinant can be greatly simplified. It is convenient, for this computation, to work with four-component Dirac fermions. Then

$$\begin{aligned} \det(\not{D}) &= \det(\not{D}\,\not{D})^{1/2} \\ &= \det\left(D^2 + \frac{1}{2} D_\mu D_\nu [\gamma^\mu, \gamma^\nu]\right) \\ &= \det(D^2 + \mathcal{F}^{\mu\nu}\mathcal{J}_{\mu\nu}). \end{aligned} \tag{3.50}$$

Here $\mathcal{F}^{\mu\nu}$ is the field strength associated with $\mathcal{A}$ (we have used the connection of the field strength and the commutator of covariant derivatives, Eq. (2.14)), and $\mathcal{J}_{\mu\nu}$ is the generator of Lorentz transformations in the fermion representation.

What is interesting is that we can write the gauge boson determinant, in the background field gauge, in a similar fashion.[1] With a little algebra, the gauge part of the action can be shown to be

$$\mathcal{L}_{\text{gauge}} = -\frac{1}{4g^2}\left(\text{Tr}\mathcal{F}^{a\ 2}_{\mu\nu} - 2g^2 a^a_\mu D^2 a^{\mu\ a} - 2a^a_\mu f^{abc}\mathcal{F}^{b\mu\nu} a^c_\nu\right). \tag{3.51}$$

[1] The details of these computations are outlined in the exercises. Here we are following closely the presentation of the text by Peskin and Schroeder.

Here we have used the A^a_μ notation, in order to be completely explicit about the gauge indices. Recalling the form of the Lorentz generators for the vector representation:

$$(\mathcal{J}^{\rho\sigma})_{\alpha\beta} = i\left(\delta^\rho_\alpha \delta^\sigma_\beta - \delta^\sigma_\alpha \delta^\rho_\beta\right), \tag{3.52}$$

we see that this object has the same formal structure as the fermion action:

$$\mathcal{L}_{\text{gauge}} = -\frac{1}{2g^2}\left[a^a_\mu \left((-D^2)^{ac} g^{\mu\nu} + 2\left(\frac{1}{2}\mathcal{F}^b_{\rho\sigma}\mathcal{J}^{\rho\sigma}\right)^{\mu\nu}\left(t^b_G\right)^{ac}\right) a^c_\nu\right]. \tag{3.53}$$

Finally, the Faddeev–Popov Lagrangian is just:

$$\mathcal{L}_c = \bar{c}^a[-(D^2)^{ab}]c^b. \tag{3.54}$$

Since the ghost fields are Lorentz scalars, this Lagrangian has the same form as the others. We need, then, to evaluate a product of determinants of the form:

$$\det\left(-D^2 + 2\left(\frac{1}{2}\mathcal{F}^b_{\rho\sigma}\mathcal{J}^{\rho\sigma}\right)t^b\right) \tag{3.55}$$

with t and $\mathcal{J}$ the generators appropriate to the representation.

The term in parentheses can be written:

$$\Delta_{r,j} = -\partial^2 + \Delta^{(1)} + \Delta^{(2)} + \Delta^{(\mathcal{J})} \tag{3.56}$$

with

$$\begin{aligned}
\Delta^{(1)} &= i\left[\partial^\mu A^a_\mu t^a + \mathcal{A}^a_\mu t^a \partial^\mu\right] \\
\Delta^{(2)} &= \mathcal{A}^{a\mu} t^a \mathcal{A}^b_\mu t^b. \\
\Delta^{(\mathcal{J})} &= 2\left(\frac{1}{2}\mathcal{F}^b_{\rho\sigma}\mathcal{J}^{\rho\sigma}\right)t^b.
\end{aligned} \tag{3.57}$$

The action we seek is the log of the determinant. We are interested in this action expanded to second order in $\mathcal{A}$ and second order in ∂^2:

$$\begin{aligned}
\ln\det(\Delta_{r,j}) = \ln\det(-\partial^2) + \text{tr}\Big[&(-\partial^2)^{-1}\left(\Delta^{(1)} + \Delta^{(2)} + \Delta^{(\mathcal{J})}\right) \\
&- \frac{1}{2}\left((-\partial^2)^{-1}\Delta^{(1)}(-\partial^2)^{-1}\Delta^{(1)}\right)\Big],
\end{aligned} \tag{3.58}$$

where $1/(-\partial^2)$ is the propagator for a scalar field. So this has the structure of a set of one-loop diagrams in a scalar field theory. Since we are working to quadratic order, we can take the $\mathcal{A}$ to carry momentum k. The term involving two powers of $\Delta^{(1)}$ is in some ways the most complicated to evaluate. Note that the trace is a trace in coordinate space and on the gauge and Lorentz indices. In momentum space, the

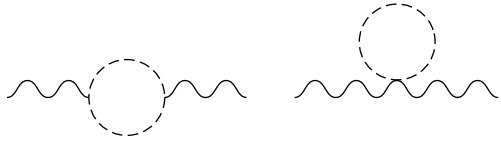

Fig. 3.4. The background field calculation has the structure of scalar electrodynamics.

space-time trace is is just an integral over momenta. We take all of the momenta to be Euclidean. So the result is given, in momentum space, by:

$$\frac{1}{2}\int \frac{d^d k}{(2\pi)^d}\mathcal{A}^a_\mu(k)\mathcal{A}^b_\nu(-k)\int \frac{d^d p}{(2\pi)^d}\mathrm{tr}\frac{1}{p^2}(2p+k)^\mu t^a \frac{1}{(p+k)^2}(2p+k)^\nu t^b. \tag{3.59}$$

This has precisely the structure of one of the vacuum polarization diagrams of scalar electrodynamics (see Fig. 3.4). The other arises from the $\Delta^{(2)}$. Combining, and performing the integral by dimensional regularization gives:

$$\frac{1}{2}\int \frac{d^d k}{(2\pi)^d}\mathcal{A}^a_\mu(-k)\mathcal{A}^b_\nu(k)(k^2 g^{\mu\nu}-k^\mu k^\nu)\left[\frac{C(r)d(j)}{3(4\pi)^2}\Gamma\left(2-\frac{d}{2}\right)(k^2)^{2-d/2}\right]. \tag{3.60}$$

where $C(r)$ are the Casimirs we have encountered previously. The quantities $C(j)$ are similar quantities for the Lorentz group; $C(j)=0$ for scalars, 1 for Dirac spinors and 2 for 4-vectors. To quadratic order in the external fields, the transverse terms above are just $(\mathcal{F}^{\mu\nu})^2$.

The piece involving $\Delta^{(\mathcal{J})}$ is even simpler to evaluate, since the needed factors of momentum (derivatives) are already included in $\mathcal{F}$. The rest is bookkeeping; the action has the form:

$$\mathcal{L}_{\text{eff}} = -\frac{1}{4}\left(\frac{1}{g^2}+\frac{1}{2}\left(C_G - C_c - \frac{n_f}{2}C_{n_f}\right)\right)F^2_{\mu\nu} \tag{3.61}$$

where

$$C_i = c_i \frac{1}{16\pi^2}\left(\frac{2}{\epsilon}-\ln(k^2)\right) \quad c_G=-20/3; c_c=3; c_{n_f}=-1/3. \tag{3.62}$$

This gives precisely Eq. (3.18).

3.6 The strong interactions and dimensional transmutation

In QCD, the only parameters, classically, with dimensions of mass are the quark masses. In a world with just two light quarks, u and d, we wouldn't expect the properties of hadrons to be different from the observed properties of the non-strange hadrons. But the masses of the up and down quarks are quite small, too small, as we

will see, to account for the masses of the non-strange hadrons, such as the proton and neutron. In other words, in the limit of zero quark mass, the hadrons would not become massless. How can a mass arise in a theory with no classical mass parameters?

While, classically, QCD is scale invariant, this is not true quantum mechanically. We have seen that we must specify the value of the gauge coupling at a particular energy scale; in the language we have used up to now, the theory is specified by giving the Lagrangian associated with a particular cutoff scale. If we change this scale, we have to change the values of the parameters, and physical quantities, such as the proton mass, should be unaffected. Using our experience with the renormalization group, we can write a differential equation which expresses how such a mass depends on g and μ so that the mass is independent of which scale we choose for our Lagrangian:

$$\left[\mu\frac{\partial}{\partial\mu}+\beta(g)\frac{\partial}{\partial g}\right]m_{\mathrm{p}}=0. \tag{3.63}$$

We know the solution of this equation:

$$m_{\mathrm{p}}=C\mu e^{-\int\frac{dg'}{\beta(g')}}. \tag{3.64}$$

To lowest order in the coupling,

$$m_{\mathrm{p}}=C\mu e^{-\frac{8\pi^2}{b_0 g^2}}. \tag{3.65}$$

This phenomenon, that a physical mass scale can appear as a result of the need to introduce a cutoff in the quantum theory is called *dimensional transmutation*. In the next section, we will discuss this phenomenon in lattice gauge theory. Later, we will describe a two-dimensional model in which we can do a simple computation which exhibits the dynamical appearance of a mass scale.

3.7 Confinement and lattice gauge theory

The fact that QCD becomes weakly coupled at high-momentum transfers has allowed rigorous comparison with experiment. Despite the fact that the variation of the coupling is only logarithmic, experiments are sufficiently sensitive, and have covered a sufficiently broad range of q^2, that such comparisons are possible. Still, many of the most interesting questions of hadronic physics – and some of the most interesting challenges of quantum field theory – are problems of low-momentum transfer. Here one encounters the flip side of asymptotic freedom: at large distances, the theory is necessarily strongly coupled and perturbative methods are not useful. It is, perhaps, frustrating that we cannot compute the masses of the low-lying hadrons in a fashion analogous to the calculation of the properties of simple atoms.

Perhaps even more disturbing is that we cannot give a simple argument that quarks are confined, or that QCD exhibits a mass gap. To deal with these questions, we can first ask a somewhat naive question: what can we say about the path integral, or for that matter the Hamiltonian, in the limit that the coupling constant becomes very large? This question is naive in that the coupling constant is not really a parameter of this theory. It is a function of scale, and the important scale for binding hadrons is the scale where the coupling becomes of order one. But let's consider the problem anyway. Start with a pure gauge theory, i.e. a theory without fermions or scalars. Consider, first, the path integral. To extract the spectrum, it should be adequate to consider the Euclidean version:

$$Z = \int [dA_\mu] \exp\left(-\frac{1}{4g^2} F_{\mu\nu}^2\right). \tag{3.66}$$

Let's contrast the weak and strong coupling limits of this expression. At weak coupling, $1/g^2$ is large, so fluctuations are highly damped; we might expect the action to be controlled by stationary points. The simplest such stationary point is just the one where $F_{\mu\nu} = 0$, and this is the basis of perturbation theory. Later we will see that there are other interesting stationary points – classical solutions of the Euclidean equations.

Now consider strong coupling. As $g \to \infty$, the action vanishes – there is no damping of quantum fluctuations. It is not obvious how one can develop any sort of approximation scheme. We can consider this problem, alternatively, from a Hamiltonian point of view. A convenient gauge for this purpose is the gauge $A_0 = 0$. In this gauge, Gauss's law is a constraint that must be imposed on states. As we will discuss shortly, Gauss's law is (almost) equivalent to the condition that the quantum states must be invariant under time-independent gauge transformations. In $A^0 = 0$ gauge, the canonical momenta are very simple:

$$\Pi^i = \frac{\partial \mathcal{L}}{\partial \dot{A}^i} = -\frac{1}{g^2} E^i. \tag{3.67}$$

So the Hamiltonian is:

$$\mathcal{H} = \frac{g^2}{2} \vec{\Pi}^2 + \frac{1}{2} \vec{B}^2. \tag{3.68}$$

In the limit $g^2 \to \infty$, the magnetic terms are unimportant, and the Π^2 terms dominate. So we should somehow work, in lowest order, with states which are eigenstates of $\vec{E}$. In any approach which respects even rotational covariance, it is unclear how to proceed.

The solution to both dilemmas is to replace the continuum of space-time with a discrete lattice of points. In the Lagrangian approach, one introduces a space-time lattice. In the Hamiltonian approach, one keeps time continuous, but makes

space discrete. Clearly there is a large price for such a move: one gives up Lorentz invariance – even rotational invariance. At best, Lorentz invariance is something which one can hope to recover in the limit that the lattice spacing is small compared to relevant physical distances. There are several rewards, however.

(1) One has a complete definition of the theory which does not rely on perturbation theory.
(2) The lattice, at strong coupling, gives a simple model of confinement.
(3) One obtains a precise procedure in which to calculate properties of hadrons. With large enough computing power, one can in principle calculate properties of low-lying hadrons with arbitrary precision.

There are other difficulties. Not only is rotational symmetry lost, but other approximate symmetries – particularly chiral symmetries – are complicated. But over time, combining ingenuity and growing computer power, there has been enormous progress in numerical lattice computations. Lattice gauge theory has developed into a highly specialized field of its own, and we will not do justice to it here. However, given the importance of field theory – often strongly coupled field theories – not only to our understanding of QCD but to any understanding of physics beyond the Standard Model, it is worthwhile to briefly introduce the subject here.

3.7.1 Wilson's formulation of lattice gauge theory

In introducing a lattice, the hope is that, as one takes the lattice spacing, a, small, one will recover Lorentz invariance. A little thought is required to understand what is meant by *small*. The only scale in the problem is the lattice spacing. But there will be another important parameter: the gauge coupling. The value of this coupling, we might expect, should be thought of as the QCD coupling at scale a. So taking small lattice spacing, physically, means taking the gauge coupling weak. At small lattice spacing, short-distance Green's functions will be well-approximated by their perturbative expansions. On the other hand, the smaller the lattice, the more numerical power required to compute the physically interesting, long-distance quantities.

There is one symmetry which one might hope to preserve as one introduces a space-time lattice: gauge invariance. Without it, there are many sorts of operators which could appear in the continuum limit, and recovering the theory of interest is likely to be very complicated. Wilson pointed out that there is a natural set of variables to work with, the Wilson lines. Consider, first, a $U(1)$ gauge theory. Under a gauge transformation, $A_\mu(x) \to A_\mu + ig(x)\partial_\mu g^\dagger(x)$, where $g(x) = e^{i\alpha(x)}$, the object

$$U(x_1, x_2) = \exp\left(i \int_{x_1}^{x_2} dx_\mu A^\mu\right),$$

transforms as:

$$U(x_1, x_2) \to g(x_1)U(x_1, x_2)g^\dagger(x_2). \tag{3.69}$$

So, for example, for a charged fermion field, $\psi(x)$, transforming as $\psi(x) \to g(x)\psi(x)$, a gauge-invariant operator is:

$$\psi^\dagger(x_1)U(x_1, x_2)\psi(x_2). \tag{3.70}$$

From gauge fields alone, one can construct an even simpler gauge-invariant object, a Wilson line beginning and ending at some point x:

$$U(x, x) = e^{i \oint_C dx^\mu A_\mu}, \tag{3.71}$$

where U is called a Wilson loop.

These objects have a simple generalization in non-Abelian gauge theories. Using the matrix form for A_μ, the main issue is one of ordering. The required ordering prescription is path ordering, P:

$$U(x_1, x_2) = P\ e^{i \int_{x_1}^{x_2} dx^\mu A_\mu}. \tag{3.72}$$

It is not hard to show that the transformation law of the Abelian case generalizes to the non-Abelian case:

$$U(x_1, x_2) \to g(x_1)U(x_1, x_2)g^\dagger(x_2). \tag{3.73}$$

To see this, note, first, that path ordering is like time ordering, so if s is the parameter of the path, U satisfies:

$$\frac{d}{ds}U(x_1(s), x_2) = \left(ig\frac{dx^\mu}{ds}A_\mu(x_1(s))\right)U(x_1(s)x_2), \tag{3.74}$$

or, more elegantly,

$$\frac{dx_1^\mu}{ds}D_\mu U(x_1, x_2) = 0. \tag{3.75}$$

Now suppose $U(x_1, x_2)$ satisfies the transformation law (Eq. (3.73)). Then it is straightforward to check, from Eq. (3.74), that $U(x_1 + dx_1, x_2)$ satisfies the correct equation. Since U satisfies a first-order differential equation, this is enough.

Again, the integral around a closed loop, C, is gauge invariant, provided one now takes the trace:

$$U(x_1, x_1) = \mathrm{Tr}\, e^{i \oint_C dx^\mu A_\mu}. \tag{3.76}$$

Wilson used these objects to construct a discretized version of the usual path integral. Take the lattice to be a simple hypercube, with points $x^\mu = an^\mu$, where n^μ is a vector of integers; a is called the lattice spacing. At any point, x, one

can construct a simple Wilson line, $U(x)_{\mu\nu}$, known as a plaquette. This is just the product of Wilson lines around a unit square. Letting n_μ denote a unit vector in the μ direction, we denote the Wilson line $U(x, x + an^\mu)$ by $U(x)_\mu$. These are the basic variables; as they are associated with the lines linking two lattice points, these are called "link variables." Then the Wilson loops about each plaquette are denoted:

$$U(x)_{\mu\nu} = U(x)_\mu U(x + an^\mu)_\nu U(x + an^\mu + an^\nu)_{-\mu} U(x + an^\nu)_{-\nu}. \tag{3.77}$$

In the non-Abelian case, a trace is understood. For small a, in the Abelian case, it is easy to expand $U_{\mu\nu}$ in powers of a, and show that:

$$U(x)_{\mu\nu} = e^{ia^2 F_{\mu\nu}(x)}. \tag{3.78}$$

So we can write an action, which in the limit of small lattice spacing goes over to the Yang–Mills action:

$$S_{\text{wilson}} = \frac{1}{4g^2} \sum_{\vec{x},\mu,\nu} U(x)_{\mu,\nu} \tag{3.79}$$

In the non-Abelian case, this same expression holds, except with 4 replaced by 2, and a trace over the U matrices.

How might we investigate the question of confinement with this action? Here, Wilson also made a proposal. Consider the amplitude for a process in which a very heavy (infinitely heavy) quark–antiquark pair are produced in the far past, separated by a distance R, and allowed to propagate for a long time, T, after which they annihilate. In Minkowski space, this amplitude would be given by:

$$\langle f | e^{-iHT} | i \rangle. \tag{3.80}$$

If we transform to Euclidean space, and insert a complete set of states, for each state we have a factor $\exp(-E_n T)$. As $T \to \infty$, this becomes $e^{-E_0 T}$, where E_0 is the ground state of the system with two infinitely massive quarks, separated by a distance R – what we would naturally identify with the potential of the quark–antiquark system.

In the path integral, this expectation value is precisely the Wilson loop, U_P, where P is the path from the point of production to the point of annihilation and back. If the quarks only experience a Coulomb force, one expects the Wilson loop to behave as

$$\langle U_P \rangle \propto e^{-\alpha T/R} \tag{3.81}$$

for a constant α. In other words, the exponential behaves as the *perimeter* of the loop. If the quarks are confined, with a linear confining force, the exponential behaves as e^{-bTR}, i.e. like the *area* of the loop. So Wilson proposed to measure the

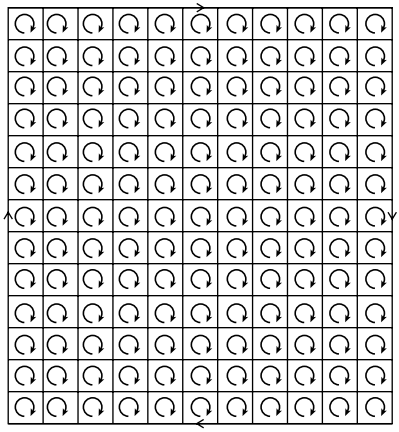

Fig. 3.5. Leading non-vanishing contribution to the Wilson loop in strong-coupling lattice gauge theory.

expectation value of the Wilson loop, and determine whether it obeyed a perimeter or area law.

In strong coupling, it is a simple matter to do the computation in the lattice gauge theory. We are interested in

$$\int \prod dU(x)_{\mu\nu} e^{-S_{\text{lattice}} + i \prod_{\text{P}} U}. \tag{3.82}$$

We can evaluate this by expanding the exponent in powers of $1/g^2$. Because

$$\int dU_\mu U_\mu = 0 \quad \int dU_\mu U_\mu U_\mu^\dagger = c \tag{3.83}$$

(you can check this easily in the Abelian case), in order to obtain a non-vanishing result, we need to tile the path with plaquettes, as indicated in Fig. 3.5. So the result is exponential in the area,

$$\langle U_P \rangle = (\text{const}/g^2)^A, \tag{3.84}$$

and the force law is

$$V(R) = \text{const} \frac{g^2}{a^2} R. \tag{3.85}$$

This is not a proof of confinement in QCD. First note that this result also holds in the strong-coupling limit of an *Abelian* gauge theory. This is possible because even the pure gauge Abelian lattice theory is an interacting theory. From this we learn that the strong-coupling behavior of a lattice theory can be very different than the weak coupling behavior. In lattice QED, there is a phase transition (a discontinuous change of behavior) between the strong- and weak-coupling phases. In the case of

an asymptotically free theory like QCD, we expect a close correspondence between coupling and lattice size. To describe, say, a proton, we would like to use a lattice on which the spacing a is much smaller than the QCD scale. In this case, the coupling we should use in the lattice theory is small. It is then not at all obvious that the strong-coupling result is applicable. At present, the issue can only be settled by evaluating the lattice path integral numerically. In principle, since the lattice theory reduces space-time to a finite number of points, the required path integral is just an ordinary integral, albeit with a huge number of dimensions. For example, if we have a 10^4 lattice, with of order 10^4 links (each 3×3 matrices), and quarks at each site, it is clear that a straightforward numerical evaluation involves an exponentially large number of operations. In practice, it is necessary to use Monte Carlo (statistical sampling) methods to evaluate the integrals. These techniques are now sufficiently powerful to convincingly demonstrate an area law at weak coupling. The constant in the area law – the coefficient of the linear term in the quark–antiquark potential, is a dimensionful parameter. It must be a renormalization-group invariant. As a result, it must take the form:

$$T = ca^{-2}e^{-\int \frac{dg'}{\beta g'}}. \tag{3.86}$$

At weak coupling, we know the form of the beta function, so we know how T should behave as we vary the lattice spacing and coupling. Results of numerical studies are in good agreement with these expressions. There has also been great progress in computing the low-lying hadron spectrum and certain weak decay matrix elements.

It is interesting to see how the strong-coupling result arises from a Hamiltonian viewpoint. To simplify the computation, we consider a $U(1)$ gauge theory. In the Hamiltonian approach, the basic dynamical variables are the matrices U_i, associated with the spatial directions. There is also A_0. As in continuum field theory, we can choose the gauge $A_0 = 0$. In this gauge, in the continuum, the dynamical variables are A_i, and their conjugate momenta E_i; on the lattice, the conjugate momenta to the U_is are the E_is. The Hamiltonian has the form:

$$H = \sum g^2 \vec{\Pi}(x)^2/a + \frac{1}{4g^2} \sum U_{ij}(\vec{x})a^{-1}. \tag{3.87}$$

The U_is are compact variables, so the $\Pi(x)$s, at each point, are like angular momenta. At strong coupling, this is a system of decoupled rotors. The ground state of the system has vanishing value of these angular momenta.

Now introduce a heavy quark–antiquark pair to the system, separated by a distance R in the z direction. In $A_0 = 0$ gauge, states must be gauge invariant (we will discuss this further when we consider instantons in the next chapter). So a candidate state has the form:

$$|\Psi\rangle = q^\dagger(0)U_z(0, R)\bar{q}^\dagger(R)|0\rangle. \tag{3.88}$$

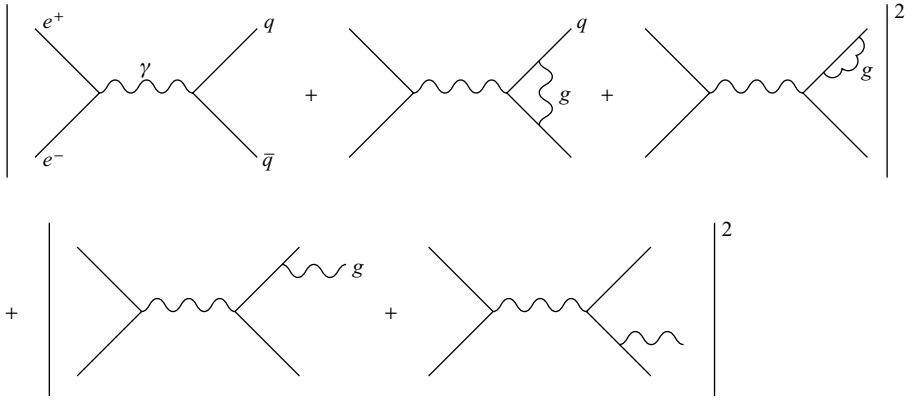

Fig. 3.6. Low-order contributions to e^+e^- annihilation.

Here

$$U_z(0, R) = U_z(0, 1)U_z(1, 2)\ldots U_z(N-1, N), \tag{3.89}$$

where $R = Na$. Now we can evaluate the expectation value of the Hamiltonian in this state. At strong coupling, we can ignore the magnetic terms. The effect of the U_z operators is to raise the "angular momentum" associated with each link by one unit (in the $U(1)$ case, $U_z(n, n+1) = e^{i\theta_{n+1}}$). So the energy of the state is just:

$$a^{-1}g^2N \tag{3.90}$$

and the potential grows linearly with separation.

3.8 Strong interaction processes at high momentum transfer

QCD has been tested, often with high precision, in a variety of processes at high momentum transfer. It is these processes to which one can apply ordinary perturbation theory. If Q^2 is the typical momentum transfer of the process, cross sections are given by a power series in $\alpha_s(q^2)$. The application of perturbation theory, however, is subtle. In accelerators, we observe hadrons; in perturbation theory, we compute the production rate for quarks and gluons. We briefly survey some of these tests in this section. In some ways, the simplest to analyze is e^+e^- annihilation, and we discuss it first. Then we turn to processes involving deep inelastic scattering of leptons by hadrons. Finally, we discuss processes involving hadrons.

3.8.1 e^+e^- Annihilation

Theoretically, perhaps the simplest process to understand is the total cross section in e^+e^- annihilation. At the level of quarks and gluons, the first few diagrams contributing to the production cross section are exhibited in Fig. 3.6. There are, in

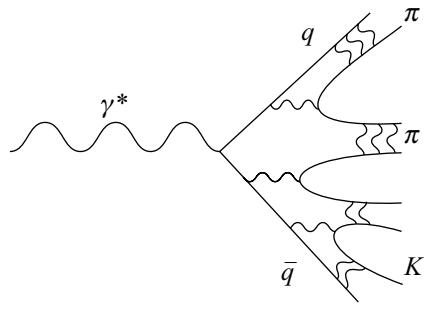

Fig. 3.7. Emission of gluons and quarks leads to formation of hadrons.

perturbation theory, a variety of final states: $q\bar{q}$, $q\bar{q}\ g$, $q\bar{q}\ g\ g$, $q\bar{q}\ q\bar{q}$, and so on. We do not understand, in any detail, how these quarks and gluons materialize as the observed hadrons. But we might imagine that this occurs as in Fig. 3.7. The initial quarks radiate gluons which can in turn radiate quark–antiquark pairs. As the cascade develops, quarks and antiquarks can pair to form mesons; qqq combinations can form baryons, and so on. In these complex processes (called "hadronization"), we can construct many relativistic invariants, and many will be small, so that perturbation theory cannot be trusted. In a sense this is good; otherwise, we would be able to show that free quarks and gluons were produced in the final states. But if we only ask about the *total* cross section, each term in the series is a function only of the center of mass energy, s. As a result, if we simply choose s for the renormalization scale, the cross section is given by a power series in $\alpha_s(s)$. One way to see this is to note that the cross section is proportional to the imaginary part of the photon vacuum polarization tensor, $\sigma(s) \propto \text{Im}\,\Pi$. One can calculate Π in Euclidean space, and then analytically continue. In the Euclidean calculation, there are no infrared divergences, so the only scales are s and the cutoff (or renormalization scale). It is convenient to consider the ratio

$$R(e^+e^- \to \text{ hadrons}) = \frac{\sigma(e^+e^- \to \text{ hadrons})}{\sigma(e^+e^- \to \mu^+\mu^-)}. \tag{3.91}$$

The lowest-order (α_s^0) contribution can be written down without any work:

$$R(e^+e^- \to \text{ hadrons}) = 3\sum Q_f^2 \tag{3.92}$$

where we have explicitly pulled out a factor of 3 for color, and the sum is over those quark flavors light enough to be produced at energy $\sqrt{s}$. So, for example, above the charm quark and below the bottom quark threshold, this would give

$$R(e^+e^- \to \text{ hadrons}) = 10/3. \tag{3.93}$$

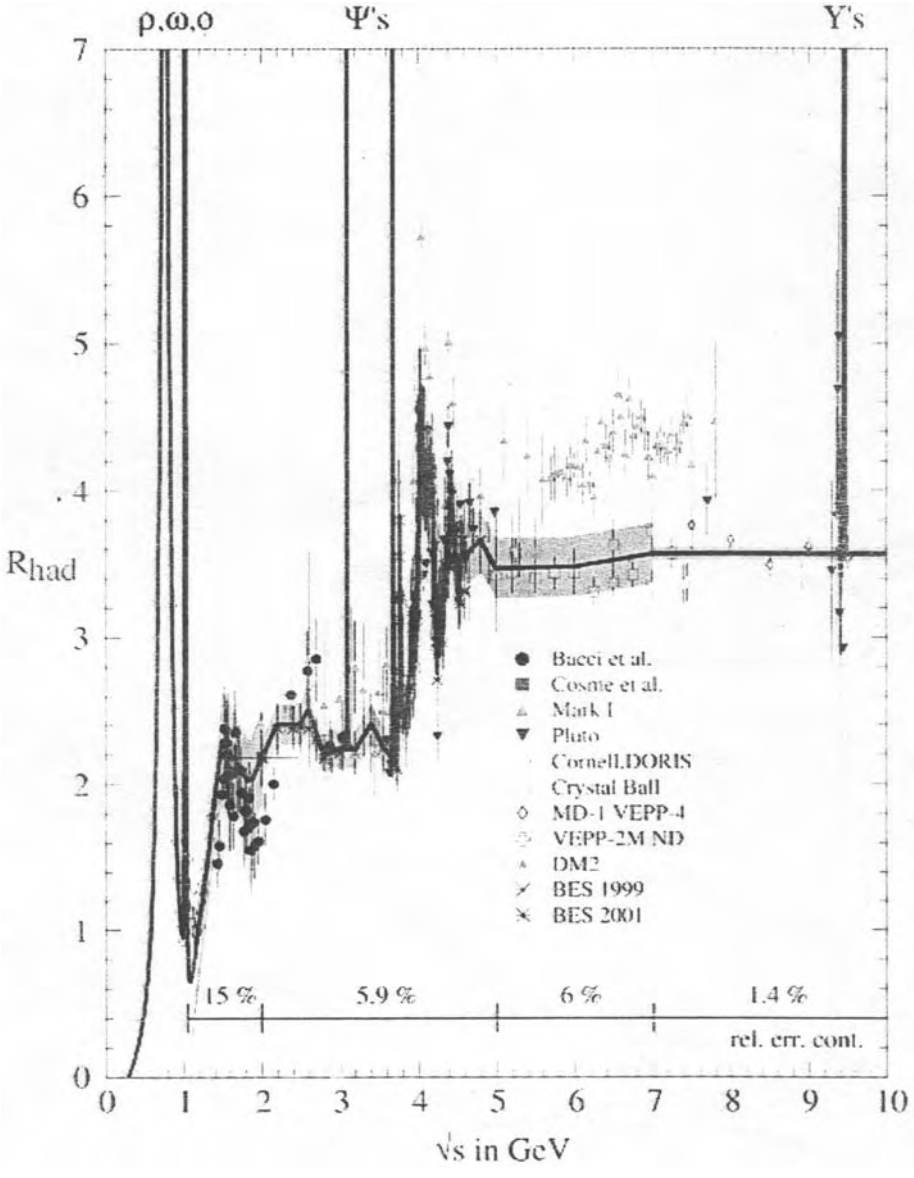

Fig. 3.8. R in e^+e^- annihilation, compared with theory. Reproduced from H. Burkhardt and B. Pietrzyk, *Phys. Lett. B* **513**, 46 (2001). Copyright 2001, with permission from Elsevier.

Before comparing with data, we should consider corrections. The cross section has been calculated through order α_s^3. Here we quote just the first two orders:

$$R(e^+e^- \to \text{ hadrons}) = 3\sum Q_f^2 \left(1 + \frac{\alpha_s}{\pi}\right). \tag{3.94}$$

This is compared with data in Fig. 3.8.

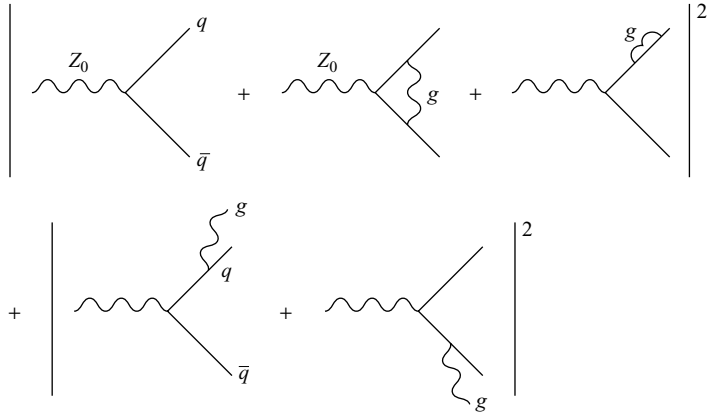

Fig. 3.9. Feynman diagrams contributing to Z decay are similar to those in e^+e^- annihilation.

Quantity	Value	Standard Model	Pull
m_t [GeV]	176.1 ± 7.4	176.96 ± 4.0	-0.1
	180.1 ± 5.4		0.6
M_W [GeV]	80.454 ± 0.059	80.390 ± 0.018	1.1
	80.412 ± 0.042		0.5
M_Z [GeV]	91.1876 ± 0.0021	91.1874 ± 0.0021	0.1
Γ_Z [GeV]	2.4952 ± 0.0023	2.4972 ± 0.0012	-0.9
Γ(had) [GeV]	1.7444 ± 0.0020	1.7435 ± 0.0011	—
Γ (inv) [MeV]	499.0 ± 1.5	501.81 ± 0.13	—
$\Gamma(\ell^+\ell^-)$ [MeV]	83.984 ± 0.086	84.024 ± 0.025	—
σ_{had} [nb]	41.341 ± 0.037	41.472 ± 0.000	1.9

Fig. 3.10. Comparison of theory and experiment for properties of the Z boson. Note close agreement at the part in $10^2 - 10^3$ level. Reprinted from *Electroweak Model and Constraints on New Physics*, Particle Data Group (2005), and S. Eidelman *et al.*, *Phys. Lett. B*, **592**, 1 (2004) (used with permission of the Particle Data Group and Elsevier).

This calculation has other applications. Perhaps the most interesting is the Z width. Z decays to hadrons involve essentially the same Feynman diagrams (Fig. 3.9), except for the different Z couplings to the quarks. Again, this is compared with experiment in Fig. 3.10.

3.8.2 Jets in e^+e^- annihilation

Much more is measured in e^+e^- annihilation than the total cross section, and clearly we would like to extract further predictions from QCD. If we are to be able

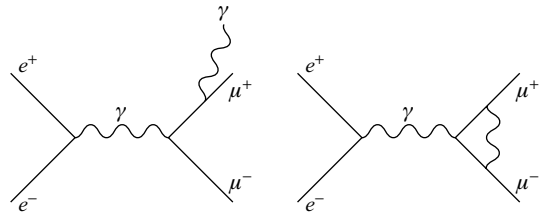

Fig. 3.11. The infrared problem.

to use perturbation theory, it is important that we limit our questions to processes for which all momentum transfers are large. It is also important that perturbation theory *fail* for some questions. After all, we know that the final states observed in accelerators contain hadrons, not quarks and gluons. If perturbation theory were good for sufficiently precise descriptions of the final state, the theory would simply be wrong.

To understand the issues, let's briefly recall some features of QED for a process like $e^+e^- \to \mu^+\mu^-$. At lowest order, one just has the production of the $\mu^+\mu^-$ pair. But in order α, one has final states with an additional photon, and loop corrections to the muon lines (also the electron/positron), as indicated in Fig. 3.11. Both the loop corrections and the total cross section for final states with a photon are infrared divergent. In QED, the resolution to this problem is *resolution*. In an experiment, one cannot detect a photon of arbitrarily low energy. So in comparing the theory with the observed cross section for $\mu^+\mu^-$ (no photon), one must allow for the possibility that a very-low-momentum photon is emitted and not detected. Including some energy resolution, ΔE, the cross sections for each possible final state are finite. If the energy is very large, one also has to keep in mind that experimental detectors cannot resolve photons sufficiently parallel to one or the other of the outgoing muons. The cross section, again, for each type of final state has large logarithms, $\ln(E/m_\mu)$. These are often called "collinear singularities" or "mass singularities." So one must allow for the finite angular resolution of real experiments. Roughly speaking, then, the radiative corrections for these processes involve

$$\delta\sigma \propto \frac{\alpha}{4\pi} \ln(E/\Delta E) \ln(\Delta\theta). \tag{3.95}$$

As one makes the energy resolution smaller, or the angular resolution smaller, perturbation theory becomes poorer. In QED, it is possible to sum these large, double-logarithmic terms.

In QCD, these same issues arise. Partial cross sections are infrared divergent. One obtains finite results if one includes an energy and angular resolution. But now the coupling is not so small as in QED, and it grows with energy. In other words, if one takes an energy resolution much smaller than the typical energy in the

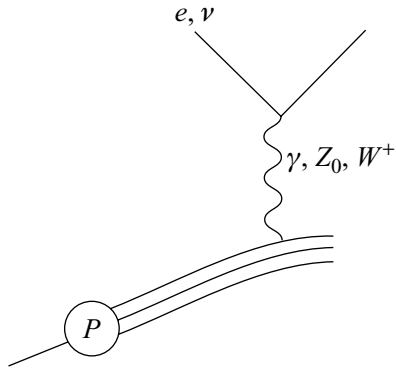

Fig. 3.12. Deep inelastic scattering of leptons off a nucleon.

process, or an angular resolution which is very small, the logarithms which appear in the perturbation expansion signal that the expansion parameter is not $\alpha_s(s)$ but something more like $\alpha_s(\Delta E)$ or $\alpha_s(\Delta\theta s)$. So perturbation theory eventually breaks down.

On the other side, if one does not make ΔE or $\Delta\theta$ too small, perturbation theory should be valid. Consider, again, e^+e^- annihilation to hadrons. One might imagine the processes which lead to the observed final states involve emission of many gluon and quark–antiquark pairs from the initial outgoing $q\bar{q}$ pair, as in Fig. 3.7. The final emissions will involve energies and momentum transfers of order the masses of pions and other light hadrons, and perturbation theory will not be useful. On the other hand, we can restrict our attention to the kinematic regime where the gluon is emitted at a large angle relative to the quark, and has a substantial energy. There are no large logarithms in this computation, nor in the computation of the $q\bar{q}$ final state. We can give a similar definition for the $q\bar{q}g$ final state. From an experimental point of view, this means that we expect to see jets of particles (or energy-momentum), reasonably collimated, and that we should be able to calculate the cross sections for emission of such jets. These calculations are similar to those of QED. Such jets are observed in e^+e^- annihilation, and their angular distribution agrees well with theoretical expectation. When first observed, these three jet events were described, appropriately, as the "discovery of the gluon."

3.8.3 Deep inelastic scattering

Deep inelastic scattering was one of the first processes to be studied theoretically in QCD. These are experiments in which a lepton is scattered at high momentum transfer from a nucleus. The lepton can be an electron, a muon, or a neutrino; the exchanged particle can be a γ, $W^\pm$ or Z (Fig. 3.12). One doesn't ask about the details

of the final hadronic state, but simply how many leptons are scattered at a given angle. Conceptually, these experiments are much like Rutherford's experiment which discovered the atomic nucleus. In much the same way, these experiments showed that nucleons contain quarks, of just the charges predicted by the quark model.

This process was attractive because one can analyze it without worrying about issues of defining jets and the like. The total cross section can be related, by unitarity, to a correlation function of two currents: the electromagnetic current, in the case of the photon; the weak currents in the case of the weak gauge bosons. The currents are space-like separated, and this separation becomes small as the momentum transfer, Q^2, becomes large. This analysis is described in many textbooks. Instead, we will adopt a different viewpoint, which allows a description of the process which generalizes to other processes involving hadrons at high momentum transfer.

Feynman and Bjorken suggested that we could view the incoming proton as a collection of quarks and gluons, which they collectively referred to as partons. They argued that one could define a probability to find a parton of type i carrying a fraction x of the proton momentum, $f_i(x)$ (and similarly for neutrons). At high momentum transfer, they suggested that the scattering of the virtual photon (or other particle) off the nucleon would actually involve the scattering of this object off one of the partons, the others being "spectators" (Fig. 3.12). In other words, the cross section for deep inelastic scattering would be given by:

$$\begin{aligned}&\sigma(e^-(k) + p(P) \to e^-(k') + X)\\ &\quad = \int d\zeta \sum_f f_f(\zeta)\sigma(e^-(k) + q(\zeta P) \to e^-(k') + q_f(p'))\end{aligned} \tag{3.96}$$

This assumption may – should – seem surprising. After all, the scattering process is described by the rules of quantum mechanics, and there should be all sorts of complicated interference effects. We will discuss this question below, but for now, suffice it to say that this picture does become correct in QCD for large momentum transfers.

For the case of the virtual photon, the cross section for the parton process can be calculated just as in QED:

$$\frac{d\sigma}{d\hat{t}}(e^-q \to e^-q) = \frac{2\pi\alpha^2 Q_f^2}{\hat{s}^2}\left[\frac{\hat{s}^2 + \hat{u}^2}{\hat{t}^2}\right]. \tag{3.97}$$

Here $\hat{s}, \hat{t}, \hat{u}$ are the kinematic invariants of the elementary parton process. For example, if we neglect the mass of the lepton and the incoming nucleon:

$$\hat{s} = 2p \cdot k = 2\zeta P \cdot k = \zeta s. \tag{3.98}$$

If the scattered electron momentum is measured, q is known, and we can relate x of the process to measured quantities. From momentum conservation:

$$(\zeta P + q)^2 = 0. \tag{3.99}$$

Or

$$q^2 + 2\zeta P \cdot q = 0. \tag{3.100}$$

Solving for ζ:

$$\zeta = x = -\frac{q^2}{2P \cdot q}. \tag{3.101}$$

It is convenient to introduce another kinematic variable,

$$y = \frac{2P \cdot q}{s} = \frac{2P \cdot q}{2P \cdot k}. \tag{3.102}$$

Then $Q^2 = q^2 = xys$, and we can write a differential cross section:

$$\frac{d^2\sigma}{dxdy}(e^- P \to e^- X) = \left(\sum_f x f_f(x) Q_f^2\right) \frac{2\pi\alpha^2 s}{Q^4}[1 + (1 - y)^2]. \tag{3.103}$$

This and related predictions were observed to hold in the first deep inelastic scattering experiments at SLAC, providing the first persuasive experimental evidence for the reality of quarks. Note, in particular, the scaling implied by these relations. For fixed y, the cross section is a function only of x.

In QCD, these notions need a crucial refinement. The distribution functions are no longer independent of Q^2:

$$f_f(x) \to f_f(x, Q^2). \tag{3.104}$$

To understand this, we return to the question: why should a probabilistic model of partons work at all in these very quantum processes? Consider, for example, the Feynman diagrams of Fig. 3.13. Clearly there are complicated interference terms when one squares the amplitude. But it turns out that, in certain gauges, the interference diagrams are suppressed, and the cross section is just given by a square of terms, as in Fig. 3.14. So one gives a probabilistic description of the process, just as Feynman and Bjorken suggested, with the distribution function the result of the sequence of interactions in the figure. These diagrams depend on Q^2. One can write integro-differential equations for these functions, the *Altareli–Parisi equations*. To explain the data, one determines these distribution functions at one value of Q^2 from experiment, and evolves them to other values. By now, these have been studied over a broad range of Q^2. Some experimental results from the HERA detector at ZEUS (DESY) are shown in Fig. 3.15.

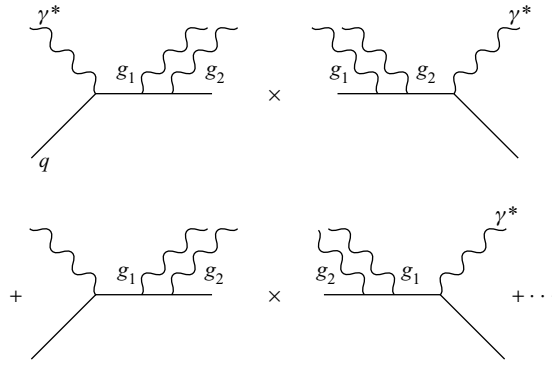

Fig. 3.13. Diagrams contributing to total rate. Diagrams to the right are complex conjugates of corresponding amplitudes on the left. The second term represents a complicated interference.

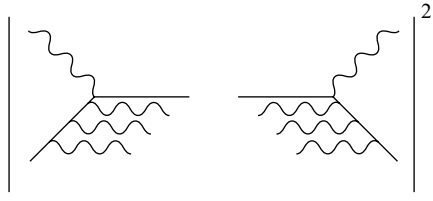

Fig. 3.14. In suitable gauges, deep inelastic scattering is dominated by absolute squares of amplitudes (interference unimportant).

3.8.4 Other high momentum processes

These ideas have been applied to other processes. The analysis which provides a diagrammatic understanding of deep inelastic scattering shows that the same structure functions are relevant to other high-momentum transfer processes, though care is required in their definitions. Examples include lepton pair production in hadronic collisions (Fig. 3.16) and jet production in hadron collisions. All have been subject to particularly stringent experimental tests.

Suggested reading

There are a number of excellent texts on the Standard Model. *An Introduction to Quantum Field Theory* by Peskin and Schroeder (1995) provides a good introduction both to the theory of weak interactions and to the strong interactions, including deep inelastic scattering, parton distributions and the like. Other excellent texts include the books by Cheng and Li (1984), Donoghue *et al*. (1992), Pokorski (2000), Weinberg (1995), Bailin and Love (1993), Cottingham and Greenwood (1998), and many others. An elegant calculation of the beta function in QCD, which uses

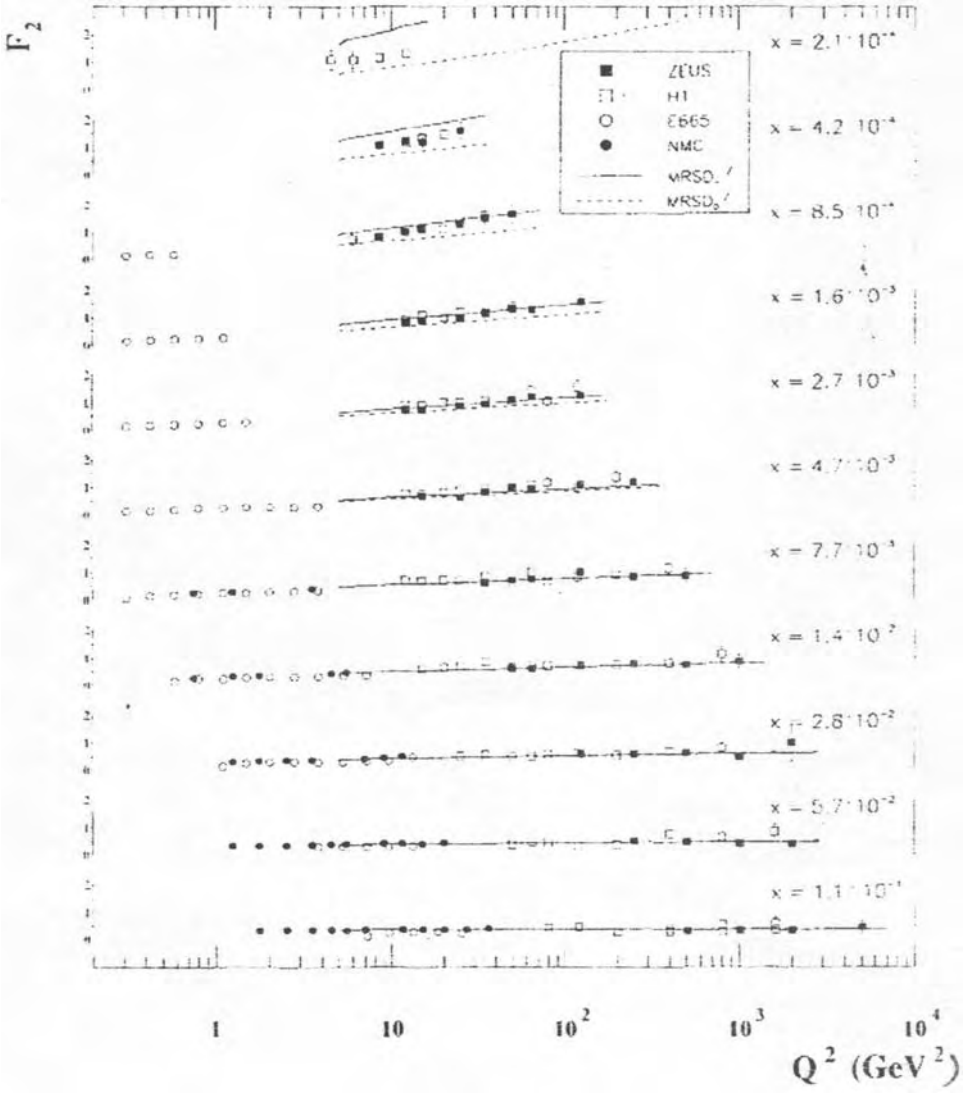

Fig. 3.15. The proton structure function F_2 as a function of Q_2 at fixed x as determined by H1 and ZEUS at DESY. From G. E. Wolf, *Nucl. Phys. Proc. Suppl*, **38**, 107 (1995).

the Wilson loop to determine the potential perturbatively, appears in the lectures of Susskind (1977). These lectures, as well as Wilson's original paper (1974) and the text of Creutz (1983), provide a good introduction to lattice gauge theory. An important subject which we did not discuss in this chapter is heavy quark physics. This is experimentally important and theoretically accessible. A good introduction is provided in the book by Manohar and Wise (2000).

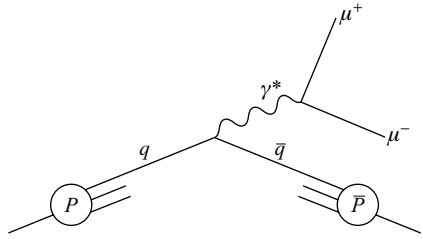

Fig. 3.16. Diagram showing $P\bar{P}$ annihilation with μ-pair production (Drell–Yan process).

Exercises

(1) Add to the Lagrangian of Eq. (2.41) a term

$$\delta\mathcal{L} = \epsilon \mathrm{Tr} M \tag{3.105}$$

for small ϵ. Show that, in the presence of ϵ, the expectation values of the $\vec{\pi}$ fields are fixed, and give a simple physical explanation. Compute the masses of the π fields directly from the Lagrangian.

(2) Verify Eqs. (2.48)–(2.56).

(3) Compute the mass of the Higgs field as a function of μ and λ. Discuss production of Higgs particles (you do not need to do detailed calculations, but indicate the relevant Feynman graphs and make at least crude estimates of cross sections) in e^+e^-, $\mu^+\mu^-$ and $P\bar{P}$ annihilation. Keep in mind that because some of the Yukawa couplings are extremely small, there may be processes generated by loop effects which are bigger than processes which arise at tree level.

(4) Using the formula for the e^+e^- cross section, determine the branching ratio for decay of the Z into hadrons:

$$B(Z \to \text{ hadrons}) = \frac{\Gamma(Z \to \text{ hadrons})}{\Gamma(Z \to \text{ all})}. \tag{3.106}$$

4

The Standard Model as an effective field theory

The Standard Model has some remarkable properties. The renormalizable terms respect a variety of symmetries, all of which are observed to hold to a high degree in Nature.

- Baryon number:

$$Q \to e^{i\frac{\alpha}{3}} Q \quad \bar{u} \to e^{-i\frac{\alpha}{3}} \bar{u} \quad \bar{d} \to e^{-i\frac{\alpha}{3}} \bar{d}. \tag{4.1}$$

- Three separate lepton numbers:

$$L_f \to e^{-i\alpha_f} L_f \quad \bar{e}_f \to e^{i\alpha_f} \bar{e}_f. \tag{4.2}$$

It is not necessary to *impose* these symmetries. They are simply consequences of gauge invariance, and the fact that there are only so many renormalizable terms one can write. These symmetries are said to be "accidental," since they don't seem to result from any deep underlying principle.

This is already a triumph. As we will see when we consider possible extensions of the Standard Model, this did not have to be. But this success raises the question: why should we impose the requirement of renormalizability?

In the early days of quantum field theory, renormalizability was sometimes presented as a sacred principle. There was a view that field theories were fundamental, and should make sense in and of themselves. Much effort was devoted to understanding whether the theories existed in the limit that the cutoff was taken to infinity.

But there was an alternative paradigm for understanding field theories, provided by Fermi's original theory of the weak interactions. In this theory, the weak interactions are described by a Lagrangian of the form:

$$\mathcal{L}_{\text{weak}} = \frac{G_f}{\sqrt{2}} J^\mu J_\mu. \tag{4.3}$$

Here the currents, J^μ, are bilinears in the fermions; they include terms like $Q\sigma^\mu T^a Q^*$. This theory, like the Standard Model, was very successful. It took some

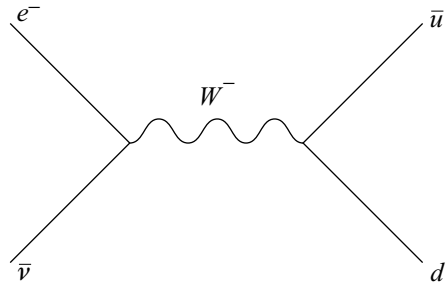

Fig. 4.1. Exchange of the massive W boson gives rise to the four-Fermi interaction.

time to actually determine the form of the currents but, for more than forty years, all experiments in weak interactions could be summarized in a Lagrangian of this form. Only as the energies of bosons in e^+e^- experiments approached the Z boson mass were deviations observed.

The four-Fermi theory is non-renormalizable. Taken seriously as a fundamental theory, it predicts violations of unitarity at TeV energy scales. But from the beginning, the theory was viewed as an *effective* field theory, valid only at low energies. When Fermi first proposed the theory, he assumed that the weak forces were caused by exchange of particles – what we now know as the W and Z bosons.

4.0.1 Integrating out the W and Z bosons

Within the Standard Model, we can derive the Fermi theory, and we can also understand the deviations. A traditional approach is to examine the Feynman diagram of Fig. 4.1. This can be understood as a contribution to a scattering amplitude, but it is best understood here as a contribution to the effective action of the quarks and leptons. The currents of the Fermi theory are just the gauge currents which describe the coupling at each vertex. The propagator, in the limit of very-small-momentum transfer, is just a constant. In coordinate space, this corresponds to a space-time δ-function – the interaction is local. The effect is just to give the four-Fermi Lagrangian. One can consider effects of small finite momentum by expanding the propagator in powers of q^2. This will give four-Fermi operators with derivatives. These are suppressed by powers of M_W and their effects are very tiny at low energies. Still, in principle, they are there, and in fact the measurement of such terms at energies a significant fraction of M_Z provided the first hints of the existence of the Z boson.

This effective action can also be derived in the path integral approach. Here we literally integrate out the heavy fields, the W and Z. In other words, for fixed values of the light fields, which we denote by ϕ, we do the path integral over W and Z,

expressing the result as an effective action for the ϕ fields:

$$\int [d\phi] e^{iS_{\text{eff}}} = \int [d\phi] \int [dW_\mu][dW_\mu^*] \Delta_{FP} e^{i\int (W_\mu^\dagger(\partial^2 + M_W^2)W^\mu + J^\mu W_\mu^\dagger + J^{\mu\dagger} W_\mu) d^4x}. \quad (4.4)$$

Here, for simplicity, we have omitted the Z particle. We have chosen the Feynman–'t Hooft gauge. The J^μ and $J^{\mu\dagger}$ are the usual weak currents. They are constructed out of the various light fields, the quarks and leptons, which we have grouped, generically, into the set of fields, ϕ. Written this way, the path integral is the most basic field theory path integral, and we are familiar with the result:

$$e^{iS_{\text{eff}}} = e^{\int d^4x d^4y J^\mu(x)\Delta(x,y)J_\mu(y)}. \quad (4.5)$$

Here $\Delta(x, y)$ denotes the propagator for a scalar of mass M_W. In the limit $M \to \infty$, this is just a δ-function (one can compute this, or see this directly from the path integral – if we neglect the derivative terms in the action, the propagator is just a constant in momentum space):

$$\Delta(x, y) = \frac{i}{M_W^2}\delta(x - y). \quad (4.6)$$

So

$$S_{\text{eff}} = \frac{1}{M_W^2} J^{\mu\dagger} J_\mu. \quad (4.7)$$

The lesson is that, up to the late 1970s, one could view QED + QCD + the Fermi theory as a perfectly acceptable theory of the interactions. The theory would have to be understood, however, as an effective theory, valid only up to an energy scale of order 100 GeV or so. Sufficiently precise experiments would require inclusion of operators of dimension higher than four. The natural scale for these operators would be the weak scale. The Fermi theory is ultraviolet divergent. These divergences would be cut off at scales of order the W boson mass.

4.0.2 What might the Standard Model come from?

As successful as the Standard Model is, and despite the fact that it is renormalizable, it is likely that, like the four-Fermi theory, it is the low-energy limit of some underlying, more "fundamental" theory. In the second half of this book, our model for this theory will be string theory. Consistent theories of strings, for reasons which are somewhat mysterious, are theories which describe general relativity and gauge interactions. Unlike field theory, string theory is a finite theory. It does not require a cutoff for its definition. In principle, all physical questions have well-defined answers within the theory. If this is the correct picture for the origin of the laws of Nature at extremely short distances, then the Standard Model is just its low-energy

limit. When we study string theory, we will understand in some detail how such a structure can emerge. For now, the main lesson we should take concerns the requirement of renormalizability: the Standard Model should be viewed as an effective theory, valid up to some energy scale, Λ. Renormalizability is not a constraint we impose upon the theory; rather, we *should* include operators of dimension five or higher with coefficients scaled by inverse powers of Λ. The question of the value of Λ is an experimental one. From the success of the Standard Model, as we will see, we know that the cutoff is large. From string theory, we might imagine that $\Lambda \approx M_{\rm p} = 1.2 \times 10^{18}$ GeV. But, as we will now describe, we have experimental evidence that there is new physics which we must include at scales well below $M_{\rm p}$. We will also see that there are theoretical reasons to believe that there should be new physics at TeV energy scales.

4.1 Lepton and baryon number violation

We have remarked that at the level of renormalizable operators, baryon and lepton number are conserved in the Standard Model. Viewed as an effective theory, however, we should include higher-dimension operators, with dimensionful couplings. We would expect such operators to arise, as in the case of the four-Fermi theory, as a result of new phenomena and interactions at very high energy scales. The coefficients of these operators would be determined by this dynamics.

There would seem, at first, to be a vast array of possibilities for operators which might be induced in the Standard Model Lagrangian. But we can organize the possible terms in two ways. First, if $M_{\rm bsm}$ is the scale of some new physics, operators of progressively higher dimension are suppressed by progressively larger powers of $M_{\rm bsm}$. Second, the most interesting, and readily detectable, operators are those which violate the symmetry of the renormalizable Lagrangian. This is already familiar in the weak interaction theory. In the Standard Model, the symmetries are precisely baryon and lepton number.

4.1.1 Dimension five: lepton number violation and neutrino mass

To proceed systematically, we should write operators of dimension five, six, and so on. At the level of dimension five, we can write several terms which violate lepton number:

$$\mathcal{L} = \frac{1}{M_{\rm bsm}} \gamma_{f,f'} \phi\phi L_f L'_f + \text{ c.c.} \tag{4.8}$$

With non-zero ϕ, these terms give rise to neutrino masses. This type of mass term is usually called a "Majorana mass." In nature, these masses are quite small. For

example, if $M_{\rm bsm} = 10^{16}$ GeV, which we will see is a plausible scale, then the neutrino masses would be of order 10^{-3} eV. In typical astrophysical and experimental situations, neutrinos are produced with energies of order MeV or larger, so it is difficult to measure these masses by studying the energy-momentum dispersion relation (very sensitive measurements of the end-point spectra beta-decay (β-decay) set limits of order electronvolts on neutrino mass). More promising are oscillation experiments, in which these operators give rise to transitions between one type of neutrino to another, similar to the phenomenon of K meson oscillations. Roughly speaking, in, say, the β-decay of a d quark, one produces the neutrino partner of the electron. However, the mass (energy) eigenstate is a linear combination of the three types of neutrino (as we will see, it is typically principally a combination of two). So experiments or observations downstream from the production point will measure processes in which neutrinos produce muons or taus.

In the past few years, persuasive evidence has emerged that the neutrinos do have non-zero masses and mixings. This comes from the study of neutrinos coming from the Sun (the "solar neutrinos") and neutrinos produced in the upper atmosphere by cosmic rays (which produce pions which subsequently decay to muons and ν_μs, whose decays in turn produce electrons, ν_μs, and $\nu_{\rm e}$s). Accelerator and reactor experiments have provided dramatic and beautiful evidence in support of this picture.

Currently, the atmospheric data is best described as a mixing of ν_μ and ν_τ. Because one is examining an oscillation, one does not know the value of the mass, but one determines:

$$\delta m^2_{\rm a} = 2.4 \times 10^{-3}\ {\rm eV}^2 \quad \theta_{\rm a} = 45^\circ \pm 10. \tag{4.9}$$

Similarly, for the solar neutrinos, which appear to be mainly mixing of ν_μ and ν_τ:

$$\delta m^2_{\rm s} = 8.2 \times 10^{-5}\ {\rm eV}^2 \quad \theta_{\rm s} = 32^\circ \pm 4. \tag{4.10}$$

It is conceivable that these masses are not described by the Lagrangian of Eq. (4.8). Instead, the masses might be "Dirac," by which one means that there might be additional degrees of freedom, which by analogy to the $\bar{e}$ fields we could label $\bar{\nu}$, with very tiny Yukawa couplings to the normal neutrinos. This would truly represent a breakdown of the Standard Model: even at low energies, we would have been missing basic degrees of freedom. But this does not seem likely. If there are singlet neutrinos, N, nothing would prevent them from gaining a "Majorana" mass:

$$\mathcal{L}_{\rm Maj} = MNN. \tag{4.11}$$

As for the leptons and quarks, there would also be a coupling of ν to N. There would now be a *mass matrix* for the neutrinos, involving both N and ν. For simplicity,

consider the case of just one generation. Then this matrix would have the form:

$$M_\nu = \begin{pmatrix} M & yv \\ yv & 0 \end{pmatrix}. \tag{4.12}$$

Such a matrix has one large eigenvalue, of order M, and one small one, of order $(y^2v^2)/M$. This provides a natural way to understand the smallness of the neutrino mass; it is referred to as the "seesaw mechanism." Alternatively, we can think of integrating out the right-handed neutrino, and generating the operator of Eq. (4.8).

It seems more plausible that the observed neutrino mass is Majorana than Dirac, but this is a question that hopefully will be settled in time by experiments searching for *neutrinoless double beta-decay*, $n + n \to p + p + e^- + e^-$. If it is Majorana, this suggests that there is another scale in physics, well below the Planck scale. For even if the new Yukawa couplings are of order one, the neutrino mass is of order

$$m_\nu = 10^{-5}\ \text{eV}(M_\text{p}/\Lambda). \tag{4.13}$$

If the Yukawas are small, as are many of the quark Yukawa couplings, the scale can be much larger.

4.1.2 Other symmetry-breaking dimension-five operators

There is another class of symmetry-violating dimension-five operators which can appear in the effective Lagrangian. These are electric and magnetic dipole moment operators. For example, the operator

$$\mathcal{L}_{\mu e} = \frac{e}{M_\text{bsm}} F_{\mu\nu}\bar{\mu}\sigma^{\mu\nu}\mu \tag{4.14}$$

(we are using a four-component notation here) would lead to processes violating muon number conservation, particularly the decay of the muon to an electron and a photon. There are stringent experimental limits on such processes, for example:

$$\text{BR}(\mu \to e\,\gamma) < 1.2 \times 10^{-11}. \tag{4.15}$$

Other operators of this type include operators which would generate lepton violating τ decays, on which the limits are far less stringent.

In the Standard Model, CP is an approximate symmetry. We have explained that three generations of quarks are required to violate CP within the Standard Model. So amplitudes which violate CP must involve all three generations, and are typically highly suppressed. From an effective Lagrangian viewpoint, if we integrate out the W and Z bosons, the operators which violate CP are dimension six, and typically have coefficients suppressed by quark masses and mixing angles, as well as loop factors. As a result, new physics at relatively modest scales has the potential for

dramatic effects. Electric dipole moment operators for quarks or leptons would arise from operators of the form:

$$\mathcal{L}_{\rm d} = \frac{em_q}{M_{\rm bsm}^2}\tilde{F}_{\mu\nu}\bar{q}\sigma^{\mu\nu}q + \text{ c.c.}, \tag{4.16}$$

where

$$\tilde{F}_{\mu\nu} = \frac{1}{2}\epsilon_{\mu\nu\rho\sigma}F^{\rho\sigma} \tag{4.17}$$

The presence of the ϵ symbol is the signal of CP violation, as the reader can check. In the non-relativistic limit, this is $\vec{\sigma}\cdot\vec{E}$. These would lead, for example, to a neutron electric dipole moment of order

$$d_{\rm n} = \frac{e}{M_{\rm bsm}}. \tag{4.18}$$

Searches for such dipole moments set limits $d_{\rm n} < 10^{-25}$ e-cm. So, unless there is some source of suppression, $M_{\rm bsm}$, in CP violating processes, is larger than about 10^2 TeV.

4.1.3 Irrelevant operators and high-precision experiments

There are a number of dimension-five operators on which it is possible to set somewhat less stringent limits, and one case in which there is a possible discrepancy. Corrections to the muon magnetic moment could arise from:

$$\mathcal{L}_{g-2} = \frac{e}{M_{\rm bsm}}F_{\mu\nu}\bar{\mu}\sigma^{\mu\nu}\mu + \text{ c.c.}, \tag{4.19}$$

where $F_{\mu\nu}$ is the electromagnetic field (in terms of the fundamental $SU(2)$ and $U(1)$ fields, one can write similar gauge-invariant combinations which reduce to this at low energies). The muon magnetic moment is measured to extremely high precision, and its Standard Model contribution is calculated with comparable precision. As of this writing, there is a 2.6 σ discrepancy between the two. Whether this reflects new physics or not is uncertain. We will encounter one candidate for this physics when we discuss supersymmetry.

There are other operators on which we can set TeV-scale limits. The success of QCD in describing jet physics allows one to constrain four quark operators which would give rise to a hard component in the scattering amplitude. Such operators might arise, for example, if quarks were composite. Constraints on flavor-changing processes provide tight constraints on a variety of operators. Operators like

$$\mathcal{L}_{\rm fc} = \frac{1}{M_{\rm bsm}^2}s\sigma^{\mu}d^{*}s\sigma_{\mu}d^{*} \tag{4.20}$$

(where we have switched to a two-component notation) would contribute to $K - \bar{K}$ mixing and other processes. This constrains $M_{\rm bsm}$ to be larger than 100 TeV or so. Any new physics at the TeV scale must explain why this operator is so severely suppressed.

4.1.4 Dimension-six operators: proton decay

Proceeding to dimension six, we can write numerous terms which violate baryon number, as well as additional lepton-number violating interactions:

$$\mathcal{L}_{\rm bv} = \frac{1}{M_{\rm bsm}^2} Q\sigma^\mu \bar{u}^* L\sigma_\mu \bar{d}^* + \cdots . \tag{4.21}$$

This can lead to processes such as $p \to \pi e$. Experiments deep underground set limits of order 10^{33} years on this process. Correspondingly, the scale $M_{\rm bsm}$ must be larger than 10^{15} GeV.

So viewing the Standard Model as an effective field theory, we see that there are many possible non-renormalizable operators which might appear, but most have scales which are tightly constrained by experiment. One might hope – or despair – that the Standard Model will provide a complete description of nature up to scales many orders of magnitude larger than we can hope to probe in experiment.

But there are a number of reasons to think that the Standard Model is incomplete, and at least one which suggests that it will be significantly modified at scales not far above the weak scale.

4.2 Challenges for the Standard Model

The Standard Model is tremendously successful. It describes the physics of strong, weak and electromagnetic interactions with great precision to energies of order 100 GeV, or distances as small as 10^{-17} cm. It explains why baryon number and the separate lepton numbers are conserved, with only one assumption: there is no interesting new physics up to some high-energy scale.

On the other hand, the Standard Model cannot be a complete theory. The existence of neutrino mass requires at least additional states (if these masses are Dirac), and more likely some new physics at a high-energy scale which accounts for the Majorana neutrino masses. This scale is probably not larger than 10^{16} GeV, well below the Planck scale. The existence of gravity means that there is certainly something missing from the theory. The plethora of parameters – about seventeen, counting those of the minimal Higgs sector – suggests that there is a deeper structure. More directly, features of the big bang cosmology which are now well established cannot be accommodated within the Standard Model.

4.2.1 A puzzle at the renormalizable level

In the standard model there is a puzzle even at the level of dimension four operators. Consider:

$$\mathcal{L}_\theta = \theta F \tilde{F}, \tag{4.22}$$

where

$$\tilde{F}_{\mu\nu} = \frac{1}{2}\epsilon_{\mu\nu\rho\sigma} F^{\rho\sigma}. \tag{4.23}$$

We usually ignore these because classically they are inconsequential; they are total derivatives and do not modify the equations of motion. In a $U(1)$ theory, for example,

$$F\tilde{F} = 2\epsilon^{\mu\nu\rho\sigma}\partial_\mu A_\nu \partial_\rho A_\sigma = 2\partial_\mu(\epsilon^{\mu\nu\rho\sigma} A_\nu \partial_\rho A_\sigma). \tag{4.24}$$

In the next chapter, we will see that this has a non-Abelian generalization, but that despite being a total divergence, these terms have real effects at the quantum level. In QCD, they turn out to be highly constrained. From the limits on the neutron electric dipole moment, one can show that $\theta < 10^{-9}$. This is the first real puzzle we have encountered. Why such a small dimensionless number? Answering this question, as we will see, may point to new physics.

4.3 The hierarchy problem

The second very puzzling feature is the Higgs field. As of this writing, the Higgs field is the one piece of the Standard Model which has not been seen. Indeed, the structure we have postulated, a single Higgs doublet with a particular potential, might be viewed as somewhat artificial. We could have included several doublets, or perhaps tried to explain the breaking of the gauge symmetry through some more complicated dynamics. But there is a more serious question associated with fundamental scalar fields, raised long ago by Ken Wilson. This problem is often referred to as the *hierarchy problem*.

Consider, first, the one-loop corrections to the electron mass in QED. These are logarithmically divergent. In other words,

$$\delta m = a m_0 \frac{\alpha}{4\pi} \ln(\Lambda). \tag{4.25}$$

We can understand this result in simple terms. In the limit $m_0 \to 0$, the theory has an additional symmetry, a chiral symmetry, under which e and $\bar{e}$ transform by independent phases. This symmetry forbids a mass term, so the result must be linear in the (bare) mass. So, on dimensional grounds, any divergence is at most logarithmic. This actually resolves a puzzle of *classical* electrodynamics. Lorentz modeled the electron as a uniformly charged sphere of radius a. As $a \to 0$, the

electrostatic energy diverges. In modern terms, we would say that we know a is smaller than 10^{-17} cm, corresponding to a self-energy far larger than the electron mass itself. But we see that in the quantum theory, the cutoff occurs at a scale of order the electron mass, and there is no large self-energy correction.

But for scalars, there is no such symmetry, and corrections to masses are quadratically divergent. One can see this quickly for the Higgs self-coupling, which gives rise to a mass correction of the form:

$$\delta m^2 = \lambda^2 \int \frac{d^4k}{(2\pi)^4(k^2 - m^2)}. \tag{4.26}$$

If we view the Standard Model as an effective field theory, this integral should be cut off at a scale where new physics enters. We have argued that this might occur at, say, 10^{14} GeV. But in this case the correction to the Higgs mass is gigantic compared to the Higgs mass itself.

4.4 Dark matter and dark energy

In a sense, our analysis seems backwards. We began with a discussion of dimension five and six operators, operators which are irrelevant, and then turned our attention to the Higgs mass, a dimension two, relevant operator. We still have not considered the most relevant operator of all, the unit operator, with dimension zero.

In recent years, astronomers and astrophysicists have presented persuasive evidence that the energy density of the universe is largely in some unfamiliar form; about 30% some non-baryonic pressureless matter (the dark matter) and about 60% in some form with negative pressure (the dark energy). The latter might be a cosmological constant (of which more later). The former could well be some new type of weakly interacting particle. Dark matter would indicate the existence of additional, possibly quite light degrees of freedom in nature. The dark energy is totally mysterious. It could be the energy of the vacuum, the *cosmological constant*. If so, it is even more puzzling than the hierarchy problem we described before. In field theory this energy is *quartically* divergent; it is the first divergence one encounters in any quantum field theory textbook. At one loop, it is given by an expression of the form:

$$\Lambda = \sum_i (-1)^F \int \frac{d^3k}{(2\pi)^3} \frac{1}{2} \sqrt{k^2 + m_i^2} \tag{4.27}$$

where the sum is over all particle species (including spins). This is just the sum of the zero-point energies of the oscillators of each momentum. If one cuts this off, again at 10^{14} GeV, one gets a result of order

$$\Lambda = 10^{54}\ \text{GeV}^4. \tag{4.28}$$

The measured value of the dark-energy density, by contrast, is

$$\Lambda = 10^{-47}\ \mathrm{GeV}^4. \tag{4.29}$$

This wide discrepancy is probably one of the most troubling problems facing fundamental physics today.

4.5 Summary: successes and limitations of the Standard Model

Overall, we face a tension between the striking successes of the Standard Model and its limitations. On the one hand, the model successfully accounts for almost all of the phenomena observed in accelerators. On the other, it fails to account for some of the most basic phenomena of the universe: dark matter, dark energy, and the existence of gravity itself. As a theoretical structure, it also explains successfully what might be viewed as mysterious conservation laws: baryon and separate lepton numbers. But it has seventeen parameters – sixteen of which are pure numbers, with values which range "all over the map." The rest of this book explores possible solutions of these puzzles, and their implications for particle physics, astrophysics and cosmology.

Suggested reading

There are a number of excellent texts on the Standard Model. *An Introduction to Quantum Field Theory* by Peskin and Schroeder (1995) provides a good introduction both to the theory of weak interactions and to the strong interactions, including deep inelastic scattering, parton distributions and the like. Other excellent texts include the books by Cheng and Li (1984), Donoghue *et al.* (1992), Pokorski (2000), Bailin and Love (1993), and many others.

5

Anomalies, instantons and the strong CP problem

While perturbation theory is a powerful and useful tool in understanding field theories, for our exploration of physics beyond the Standard Model, an understanding of non-perturbative physics will be crucial. There are many reasons for this.

(1) One of the great mysteries of the Standard Model is non-perturbative in nature: the smallness of the θ parameter.
(2) Strongly interacting field theories will figure in many proposals to understand mysteries of the Standard Model.
(3) The interesting dynamical properties of supersymmetric theories, both those directly related to possible models of nature and those which provide insights into broad physics issues, are non-perturbative in nature.
(4) If string theory describes nature, non-perturbative effects are necessarily critical.

We have introduced lattice gauge theory, which is perhaps our only tool for doing systematic calculations in strongly coupled theories. But, as a tool, its value is quite limited. Only a small number of calculations are tractable, in practice, and the difficult numerical challenges sometimes obscure the underlying physics. Fortunately, there is a surprising amount that one can learn from symmetry considerations, semiclassical arguments and from our experimental knowledge of one strongly coupled theory, QCD. In each of these, an important role is played by the phenomena known as anomalies, and related to these a set of semiclassical field configurations known as instantons.

Usually, the term anomaly is used to refer to the quantum-mechanical violation of a symmetry which is valid classically. Instantons are finite-action solutions of the Euclidean equations of motion, typically associated with tunneling phenomena. Anomalies are crucial to understanding the decay of the π^0 in QCD. Anomalies and instantons account for the absence of a ninth light pseudoscalar meson in the hadron spectrum. Within the weak-interaction theory, anomalies and instantons lead to violation of baryon and lepton number; these effects are unimaginably

tiny at the current time, but were important in the early universe. Absence of anomalies in gauge currents is important to the consistency of theoretical structures, including both field theories and string theories. The cancellation of anomalies within the Standard Model itself is quite non-trivial. Similar constraints on possible extensions of the Standard Model will be very important. In the previous chapter, we mentioned the θ parameter of QCD. This term seems innocuous, but, due to anomalies and instantons, its potential effects are real. Because the θ term violates CP, they are also dramatic. The problem of the smallness of the θ parameter – the *strong CP problem* – strongly suggests new phenomena beyond the Standard Model, and this will be a recurring theme in this book. In this chapter, we explain how anomalies arise and some of the roles which they play. The discussion is meant to provide the reader with a good working knowledge of these subjects, but it is not encyclopedic. A guide to texts and reviews on the subject appears at the end of the chapter.

5.1 The chiral anomaly

Before considering real QCD, consider a non-Abelian gauge theory theory, with only a single flavor of quark. Before making any field redefinitions, the Lagrangian takes the form:

$$\mathcal{L} = -\frac{1}{4g^2}F_{\mu\nu}^2 + \bar{q}\,D^\mu\sigma_\mu\bar{q}^* + q\,D^\mu\sigma_\mu q^* + m\bar{q}q + m^*\bar{q}^*q^*. \tag{5.1}$$

The Lagrangian, here, is written in terms of two-component fermions (see Appendix A). The fermion mass need not be real,

$$m = |m|e^{i\theta}. \tag{5.2}$$

In this chapter, it will sometimes be convenient to work with four component fermions, and it is valuable to make contact with this language in any case. In terms of these, the mass term is:

$$\mathcal{L}_m = \mathrm{Re}(m)\,\bar{q}q + \mathrm{Im}(m)\,\bar{q}\gamma_5 q. \tag{5.3}$$

In order to bring the mass term to the conventional form, with no γ_5s, one could try to redefine the fermions; switching back to the two-component notation:

$$q \to e^{-i\theta/2}q \quad \bar{q} \to e^{-i\theta/2}\bar{q}. \tag{5.4}$$

But, in field theory, transformations of this kind are potentially fraught with difficulties because of the infinite number of degrees of freedom.

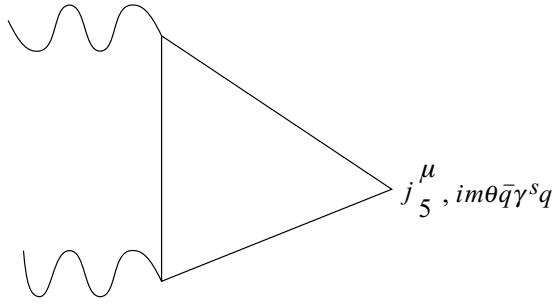

Fig. 5.1. The triangle diagram associated with the four-dimensional anomaly.

A simple calculation uncovers one of the simplest manifestations of an anomaly. Suppose, first, that m is very large, $m \to M$. In that case we want to integrate out the quarks and obtain a low-energy effective theory. To do this, we study the path integral:

$$Z = \int [dA_\mu] \int [dq][d\bar{q}] e^{iS} \tag{5.5}$$

Suppose $M = e^{i\theta}|M|$. In order to make m real, we can again make the transformations: $q \to qe^{-i\theta/2}; \bar{q} \to \bar{q}e^{-i\theta/2}$ (in four-component language, this is $q \to e^{-i\theta/2\gamma_5}q$).) The result of integrating out the quark, i.e. of performing the path integral over q and $\bar{q}$, can be written in the form:

$$Z = \int [dA_\mu] \int e^{iS_{\text{eff}}}. \tag{5.6}$$

Here S_{eff} is the effective action which describes the interactions of gluons at scales well below M.

Because the field redefinition which eliminates θ is just a change of variables in the path integral, one might expect that there can be no θ-dependence in the effective action. But this is not the case. To see this, suppose that θ is small, and instead of making the field redefinition, treat the θ term as a small perturbation by expanding the exponential. Now consider a term in the effective action with two external gauge bosons. This is obtained from the Feynman diagram in Fig. 5.1. The corresponding term in the action is given by

$$\delta\mathcal{L}_{\text{eff}} = -i\frac{\theta}{2}g^2 M \text{Tr}(T^a T^b) \int \frac{d^4p}{(2\pi)^4} \text{Tr}\gamma_5 \frac{1}{\not{p} + \not{k}_1 - M} \not{\epsilon}_1 \frac{1}{\not{p} - M} \not{\epsilon}_2 \frac{1}{\not{p} - \not{k}_2 - M}. \tag{5.7}$$

Here, the k_is are the momenta of the two photons, while the ϵs are their polarizations and a and b are the color indices of the gluons. Introducing Feynman parameters

and shifting the p integral, gives:

$$\delta\mathcal{L}_{\text{eff}} = -i\theta g^2 M \text{Tr}(T^a T^b) \int d\alpha_1 d\alpha_2 \int \frac{d^4p}{(2\pi)^4} \text{Tr}\gamma_5 (\not{p} - \alpha_1 \not{k}_1 + \alpha_2 \not{k}_2 + \not{k}_1 + M) \not{\epsilon}_1 \times (\not{p} - \alpha_1 \not{k}_1 + \alpha_2 \not{k}_2 + M) \not{\epsilon}_2 (\not{p} - \alpha_1 \not{k}_1 + \alpha_2 \not{k}_2 - \not{k}_2 + M) \Big/ \left(p^2 - M^2 + O\left(k_i^2\right)\right)^3. \tag{5.8}$$

For small k_i, we can neglect the k-dependence of the denominator. The trace in the numerator is easy to evaluate, since we can drop terms linear in p. This gives, after performing the integrals over the αs,

$$\delta\mathcal{L}_{\text{eff}} = g^2 M^2 \theta \text{Tr}(T^a T^b) \epsilon_{\mu\nu\rho\sigma} k_1^\mu k_2^\nu \epsilon_1^\rho \epsilon_2^\sigma \int \frac{d^4p}{(2\pi)^4} \frac{1}{(p^2 - M^2)^3}. \tag{5.9}$$

This corresponds to a term in the effective action, after doing the integral over p and including a combinatoric factor of two from the different ways to contract the gauge bosons:

$$\delta\mathcal{L}_{\text{eff}} = \frac{1}{32\pi^2} \theta \text{Tr}(F\tilde{F}). \tag{5.10}$$

Now why does this happen? At the level of the path integral, the transformation would seem to be a simple change of variables, and it is hard to see why this should have any effect. On the other hand, if one examines the diagram of Fig. 5.1, one sees that it contains terms which are linearly divergent, and thus it should be regulated. A simple way to regulate the diagram is to introduce a Pauli–Villars regulator, which means that one subtracts off a corresponding amplitude with some very large mass Λ. However, our expression above is independent of Λ. So the θ-dependence from the regulator fields cancels that of Eq. (5.10). This sort of behavior is characteristic of an anomaly.

Consider now the case that $m \ll \Lambda_{\text{QCD}}$. In this case, we shouldn't integrate out the quarks, but we still need to take into account the regulator diagrams. So if we redefine the fields so that the quark mass is real (γ_5-free, in the four-component description), the low-energy theory contains light quarks and the θ term of Eq. (5.10).

We can describe this in a fashion which indicates why this is referred to as an *anomaly*. For small m, the classical theory has an approximate symmetry under which

$$q \to e^{i\alpha} q \quad \bar{q} \to e^{i\alpha} \bar{q} \tag{5.11}$$

(in four-component language, $q \to e^{i\alpha\gamma_5} q$). In particular, we can define a current:

$$j_5^\mu = \bar{q}\gamma_5\gamma_\mu q \tag{5.12}$$

and, classically,

$$\partial_\mu j_5^\mu = m\bar{q}\gamma_5 q. \tag{5.13}$$

Under a transformation by an infinitesimal angle α one would expect

$$\delta L = \alpha \partial_\mu j_5^\mu = m\alpha\bar{q}\gamma_5 q. \tag{5.14}$$

But the divergence of the current contains another, m-independent term:

$$\partial_\mu j_5^\mu = m\bar{q}\gamma_5 q + \frac{1}{32\pi^2} F\tilde{F}. \tag{5.15}$$

The first term just follows from the equations of motion. To see that the second term is present, we can study a three-point function involving the current and two gauge bosons, ignoring the quark mass:

$$\Gamma^{AAj} = T\langle \partial_\mu j^{5\,\mu} A_\rho A_\sigma \rangle. \tag{5.16}$$

This is essentially the calculation we encountered above. Again, the diagram is linearly divergent and requires regularization. Let's first consider the graph without the regulator mass. The graph of Fig. 5.1 is actually two graphs, because we must include the interchange of the two external gluons. The combination is easily seen to vanish, by the sorts of manipulations one usually uses to prove Ward identities:

$$\frac{g^2}{(2\pi)^4}\int d^4p \,\mathrm{Tr}\, \not{q}\gamma_5 \frac{1}{\not{p}+\not{k}_1} \not{\epsilon}_1 \frac{1}{\not{p}} \not{\epsilon}_2 \frac{1}{\not{p}-\not{k}_2} + (1 \leftrightarrow 2). \tag{5.17}$$

Writing

$$\not{q}\gamma_5 = -\gamma_5(\not{k}_1 + \not{p}) - (\not{p} - \not{k}_2)\gamma_5 \tag{5.18}$$

and using the cyclic property of the trace, one can cancel a propagator in each term. This leaves:

$$\int d^4p \,\mathrm{Tr}\left(-\gamma_5 \not{\epsilon}_1 \frac{1}{\not{p}} \not{\epsilon}_2 \frac{1}{\not{p}-\not{k}_2} - \gamma_5 \frac{1}{\not{p}+\not{k}_1} \not{\epsilon}_1 \frac{1}{\not{p}} \not{\epsilon}_2 + (1 \leftrightarrow 2)\right). \tag{5.19}$$

Now shifting $p \to p + k_2$ in the first term, and $p \to p + k_1$ in the second, one finds a pairwise cancellation.

These manipulations, however, are not reliable. In particular, in a highly divergent expression, the shifts do not necessarily leave the result unchanged. With a Pauli–Villars regulator, the integrals are convergent and the shifts are reliable, but the regulator diagram is non-vanishing, and gives the anomaly equation above. One can see this by a direct computation, or relate it to our previous calculation, including the masses for the quarks, and noting that $\not{q}\gamma_5$, in the diagrams with massive quarks, can be replaced by $M\gamma_5$.

This anomaly can be derived in a number of other ways. One can define, for example, the current by "point splitting,"

$$j_5^\mu = \bar{q}(x + i\epsilon)e^{i\int_x^{x+\epsilon} dx^\mu A_\mu} q(x). \tag{5.20}$$

Because operators in quantum field theory are singular at short distances, the Wilson line makes a finite contribution. Expanding the exponential carefully, one recovers the same expression for the current. We will do this shortly in two dimensions, leaving the four-dimensional case for the problems. A beautiful derivation, closely related to that we have performed above, is due to Fujikawa. Here one considers the anomaly as arising from a lack of invariance of the path integral measure. One carefully evaluates the Jacobian associated with the change of variables $q \to q(1 + i\gamma_5\alpha)$, and shows that it yields the same result. We will do a calculation along these lines in a two-dimensional model shortly, leaving the four-dimensional case for the problems.

5.1.1 Applications of the anomaly in four dimensions

The anomaly has a number of important consequences for real physics.

- The π^0 decay: the divergence of the axial isospin current,

$$\left(j_5^3\right)^\mu = \bar{u}\gamma_5\gamma^\mu\bar{u} - \bar{d}\gamma_5\gamma^\mu d, \tag{5.21}$$

  has an anomaly due to electromagnetism. This gives rise to a coupling of the π^0 to two photons, and the correct computation of the lifetime was one of the early triumphs of the theory of quarks with color. The computation of the π^0 decay rate appears in the exercises.
- Anomalies in gauge currents signal an inconsistency in a theory. They mean that the gauge invariance, which is crucial to the whole structure of gauge theories (e.g. to the fact that they are simultaneously unitary and Lorentz-invariant) is lost. The absence of gauge anomalies is one of the striking ingredients of the Standard Model, and it is also crucial in extensions such as string theory.
- The anomaly, as we have indicated, accounts for the absence of a ninth axial Goldstone boson in the QCD spectrum.

5.1.2 Return to QCD

What we have just learned is that, if in our simple model above, we require that the quark masses are real, we must allow for the possible appearance, in the Lagrangian of the standard model, of the θ terms of Eq. (5.10). In the weak interactions, this term does not have physical consequences. At the level of the renormalizable terms, we have seen that the theory respects separate B and L symmetries; B, for example,

is anomalous. So if we simply redefine the quark fields by a B transformation, we can remove θ from the Lagrangian.

For the θ angles of QCD and QED, we have no such symmetry. In the case of QED, we do not really have a non-perturbative definition of the theory, and the effects of θ are hard to assess, but one might expect that, embedded in any consistent structure (like a grand unified theory (GUT) or String Theory) they will be very small, possibly zero. As we saw, $F\tilde{F}$ is a total divergence. The right-hand side of Eq. (4.24) is not gauge invariant, however, so one might imagine that it could be important. But as long as A falls off at least as fast as $1/r$ (F falls faster than $1/r^2$), the surface term behaves as $1/r^4$, and so vanishes.

In the case of non-Abelian gauge theories, the situation is more subtle. It is again true that $F\tilde{F}$ can be written as a total divergence:

$$F\tilde{F} = \partial^\mu K_\mu \quad K_\mu = \epsilon_{\mu\nu\rho\sigma}\left(A^a_\nu F^a_{\rho\sigma} - \frac{2}{3} f^{abc} A^a_\nu A^b_\rho A^c_\sigma\right). \tag{5.22}$$

But now the statement that F falls faster than $1/r^2$ does not permit an equally strong statement about A. We will see shortly that there are finite-action classical solutions for which $F \sim 1/r^4$, but $A \to 1/r$, so that the surface term cannot be neglected. These solutions are called instantons. It is because of this that θ can have real physical effects.

5.2 A two-dimensional detour

There are many questions in four dimensions which we cannot answer except with numerical lattice calculation. These include the problem of dimensional transmutation and the effects of the anomaly on the hadron spectrum. There are a class of models in two dimensions which are asymptotically free and in which one can study these questions in a controlled approximation. Two dimensions are often a poor analog for four, but for some of the issues we are facing here, the parallels are extremely close. In these two-dimensional examples, the physics is more manageable, but still rich. In four dimensions, the calculations are qualitatively similar; they are only more difficult because the Dirac algebra and the various integrals are more involved.

5.2.1 The anomaly in two dimensions

First we investigate the anomaly in the quantum electrodynamics of a massless fermion in two dimensions; this will be an important ingredient in the full analysis. The point-splitting method is particularly convenient here. Just as in four

dimensions, we write:

$$j_5^\mu = \bar{\psi}(x+\epsilon)e^{i\int_x^{x+\epsilon} A_\rho dx^\rho}\gamma^\mu\gamma^5\psi(x). \tag{5.23}$$

Naively, one can set $\epsilon = 0$ and the divergence vanishes by the equations of motion. But in quantum field theory, products of operators become singular as the operators come very close together. For very small ϵ, we can pick up the leading singularity in the product of $\psi(x+\epsilon)\psi$ by using the operator product expansion (OPE). The OPE states that the product of two operators at short distances can be written as a series of local operators of progressively higher dimension, with less and less singular coefficients. For our case, this means:

$$\bar{\psi}(x+\epsilon)\gamma^\mu\gamma^5\psi(x) = \sum \frac{c_n}{\epsilon^{1-n}}\mathcal{O}_n(x) \tag{5.24}$$

where $\mathcal{O}_n$ is an operator of dimension n. The leading term comes from the unit operator. To evaluate its coefficient, we can take the vacuum expectation value of both sides of this equation. On the left-hand side, this is just the propagator.

It is not hard to work out the fermion propagator in coordinate space in two dimensions. For simplicity, we work with space-like separations, so we can Wick-rotate to Euclidean space. Start with the scalar propagator:

$$\begin{aligned}\langle\phi(x)\phi(0)\rangle &= \int \frac{d^2p}{(2\pi)^2}\frac{1}{k^2}e^{-ip\cdot x} \\ &= \frac{1}{2\pi}\ln(|x|\mu),\end{aligned} \tag{5.25}$$

where μ is an infrared cutoff. (When we come to string theory, this propagator, with its infrared sensitivity, will play a crucial role.) Correspondingly, the fermion propagator is:

$$\langle\bar{\psi}(x+\epsilon)\psi(x)\rangle = \not{\partial}\langle\phi(x)\phi(0)\rangle = \frac{1}{2\pi}\frac{\not{\epsilon}}{\epsilon^2}. \tag{5.26}$$

Expanding the factor in the exponential to order ϵ gives

$$\partial_\mu j_5^\mu = \text{naive piece} + \frac{i}{2\pi}\partial_\mu\epsilon_\rho A^\rho \text{tr}\frac{\not{\epsilon}}{\epsilon^2}\gamma^\mu\gamma^5. \tag{5.27}$$

Taking the trace gives $\epsilon_{\mu\nu}\epsilon^\nu$; averaging ϵ over angles ($\langle\epsilon_\mu\epsilon_\nu\rangle = \frac{1}{2}\eta_{\mu\nu}\epsilon^2$), yields

$$\partial_\mu j_5^\mu = \frac{1}{2\pi}\epsilon_{\mu\nu}F^{\mu\nu}. \tag{5.28}$$

This is parallel to the situation in four dimensions. The divergence of the current is itself a total derivative:

$$\partial_\mu j_5^\mu = \frac{1}{2\pi}\epsilon_{\mu\nu}\partial^\mu A^\nu. \tag{5.29}$$

So it is possible to define a new current which is conserved:

$$J^\mu = j_5^\mu - \frac{1}{2\pi}\epsilon^\mu_\nu A^\nu \tag{5.30}$$

However, just as in the four-dimensional case, this current is not gauge invariant. There is a familiar field configuration for which A does not fall off at infinity: the field of a point charge. If one has charges, $\pm\theta$ at infinity, they give rise to a constant electric field, $F_{0i} = \pm e\theta$. So θ has a very simple interpretation in this theory.

It is easy to see that physics is periodic in θ. For $\theta > q$, it is energetically favorable to produce a pair of charges from the vacuum which shield the charge at ∞.

5.2.2 Path integral computation of the anomaly

One can also do this calculation in the path integral, following Fujikawa. The redefinition of the fields which eliminates the phase in the fermion mass matrix, from this point of view, is just a change of variables. The question is: what is the Jacobian. The Euclidean path integral is defined by expanding the fields:

$$\psi(x) = \sum a_n \psi_n(x) \tag{5.31}$$

where

$$\not{D}\psi_n(x) = \lambda_n \psi_n(x) \tag{5.32}$$

and the measure is:

$$\int \Pi da_n da_n^*. \tag{5.33}$$

Here, for normalized functions ψ_n,

$$a_n = \int d^2x \psi_n^*(x)\psi(x). \tag{5.34}$$

So, under an infinitesimal γ_5 transformation, we have:

$$\delta\psi = i\theta\gamma_5\psi \tag{5.35}$$

$$\delta a_n = i\theta \int d^2x \psi_n(x)\gamma_5\psi_m(x)a_m. \tag{5.36}$$

The required Jacobian is then:

$$\det\left(\delta_{nn'} + i\theta \int \bar{\psi}_{n'}\gamma_5\psi_n\right). \tag{5.37}$$

To evaluate this determinant, we write $\det(M) = e^{\mathrm{Tr}\log M}$. To linear order in θ, we need to evaluate:

$$\mathrm{Tr}\, i\theta\gamma_5. \tag{5.38}$$

This trace must be regularized. A simple procedure is to replace the determinant by:

$$\mathrm{Tr}\, i\theta\gamma_5 e^{-(\lambda_n^2/M^2)}. \tag{5.39}$$

At the end of the calculation, we should take $M \to \infty$. We can replace λ_n^2 by

$$\not{D}\not{D} = D^2 + \frac{1}{2}\sigma_{\mu\nu}F^{\mu\nu}. \tag{5.40}$$

Expanding in powers of $F^{\mu\nu}$, it is only necessary to work to first order (in the analogous calculation in four dimensions, it is necessary to work to second order). In other words, we expand the exponent to first order in $F^{\mu\nu}$ and replace $D^2 \to p^2$. The required trace is:

$$i\theta \int \frac{d^2p}{p^2}\mathrm{Tr}(\gamma_5\sigma_{\mu\nu})\frac{F^{\mu\nu}}{M}e^{-\frac{p^2}{M^2}}. \tag{5.41}$$

The trace here now just refers to the trace over the Dirac indices. The momentum integral is elementary, and we obtain

$$\int \Pi da_n da_n^* \to \int \Pi da_n da_n^* e^{i\frac{\theta}{2\pi}\int d^2x \epsilon_{\mu\nu}F^{\mu\nu}}. \tag{5.42}$$

Interpreting the divergence of the current as the variation of the effective Lagrangian, we see that we have recovered the anomaly equation. The anomaly in four and other dimensions can also be calculated in this way. The exercises at the end of the chapter provide more details of these computations.

5.2.3 *The* CP^N *model: an asymptotically free theory*

The model we have considered so far is not quite like QCD in at least two ways. First, there are no instantons; second, the coupling e is dimensionful. We can obtain a theory closer to QCD by considering a class of theories with dimensionless couplings, the *non-linear* σ*-models*. These are models whose fields are the coordinates of some smooth manifold. They can be, for example, the coordinates of an n-dimensional sphere. An interesting case is the CP^N model; here the CP stands for "complex projective" space. This space is described by a set of coordinates, z_i, $i = 1, \ldots, N+1$, where z_i is identified with αz_i, where α is any complex constant. Alternatively, we can define the space through the constraint:

$$\sum_i |z_i|^2 = 1; \tag{5.43}$$

where the point z_i is equivalent to $e^{i\alpha}z_i$. In the field theory, the z_is become two-dimensional fields, $z_i(x)$. To implement the first of the constraints, we can add to the action a Lagrange multiplier field, $\lambda(x)$. For the second, we observe that

the identification of points in the "target space," CP^N, must hold at every point in *ordinary space-time*, so this is a $U(1)$ gauge symmetry. So introducing a gauge field, A_μ, and the corresponding covariant derivative, we want to study the Lagrangian:

$$\mathcal{L} = \frac{1}{g^2}[|D_\mu z_i|^2 - \lambda(x)(|z_i|^2 - 1)]. \tag{5.44}$$

Note that there is no kinetic term for A_μ, so we can simply eliminate it from the action using its equations of motion. This yields

$$\mathcal{L} = \frac{1}{g^2}[|\partial_\mu z_j|^2 + |z_j^* \partial_\mu z_j|^2]. \tag{5.45}$$

It is easier to proceed, however, keeping A_μ in the action. In this case, the action is quadratic in z, and we can integrate out the z fields:

$$\begin{aligned} Z &= \int [dA][d\lambda][dz_j] \exp[-S] = \int [dA][d\lambda] \exp\left[-\int d^2x \Gamma_{\text{eff}}[A, \lambda]\right] \\ &= \int [dA][d\lambda] \exp\left[-N \operatorname{tr} \log(-D^2 - \lambda) - \frac{1}{g^2}\int d^2x \lambda\right]. \end{aligned} \tag{5.46}$$

5.2.4 The large-N limit

By itself, the result of Eq. (5.46) is still rather complicated. The fields A_μ and λ have non-linear and non-local interactions. Things become much simpler if one takes the "large N limit," $N \to \infty$ with $g^2 N$ fixed. In this case, the interactions of λ and A_μ are suppressed by powers of N. For large N, the path integral is dominated by a single field configuration, which solves

$$\frac{\delta \Gamma_{\text{eff}}}{\delta \lambda} = 0 \tag{5.47}$$

or, setting the gauge field to zero,

$$N \int \frac{d^2k}{(2\pi)^2} \frac{1}{k^2 + \lambda} = \frac{1}{g^2}. \tag{5.48}$$

The integral on the left-hand side is ultraviolet divergent. We will simply cut it off at scale M. This gives:

$$\lambda = m^2 = M \exp\left[-\frac{2\pi}{g^2 N}\right]. \tag{5.49}$$

Here, a theory which is scale invariant, classically, exhibits a mass gap. This is the phenomenon of *dimensional transmutation*. These masses are related in a renormalization-group-invariant fashion to the cutoff. So the theory is quite analogous to QCD. We can read off the leading term in the beta (β-)function from the

familiar formula:

$$m = Me^{-\int \frac{dg}{\beta(g)}} \tag{5.50}$$

so, with

$$\beta(g) = -\frac{1}{2\pi} g^3 b_0 \tag{5.51}$$

we have $b_0 = 1$.

But most important for our purposes, it is interesting to explore the question of θ-dependence. Just as in $1+1$-dimensional electrodynamics, we can introduce a θ term:

$$S_\theta = \frac{\theta}{2\pi} \int d^2x \epsilon_{\mu\nu} F^{\mu\nu}. \tag{5.52}$$

Here $F_{\mu\nu}$ can be expressed in terms of the fundamental fields z_j. As usual, this is the integral of a total divergence. But precisely as in the case of $1+1$-dimensional electrodynamics we discussed above, this term is physically important. In perturbation theory in the model, this is not entirely obvious. But using our reorganization of the theory at large N, it is. The lowest-order action for A_μ is trivial, but at one loop (order $1/N$), one generates a kinetic term for A through the vacuum polarization loop:

$$\mathcal{L}_{\text{kin}} = \frac{N}{2\pi m^2} F_{\mu\nu}^2. \tag{5.53}$$

At this order, then, the effective theory consists of the gauge field, with coupling $e^2 = 2\pi m^2/N$, and some coupling to a set of charged massive fields, z. As we have already argued, θ corresponds to a non-zero background electric field due to charges at infinity, and the theory clearly has non-trivial θ-dependence.

To this model one can add massless fermions. In this case, one has an anomalous $U(1)$ symmetry, as in QCD. There is then no θ-dependence; by redefining the fermions, $\psi \to e^{ia\theta}\psi$, one can eliminate θ. In this model, the absence of θ-dependence can be understood more physically: θ represents a charge at ∞, and it is possible to shield any such charge with massless fermions. But there is non-trivial breaking of the $U(1)$ symmetry. At low energies, one has now a theory with a fermion coupled to a dynamical $U(1)$ gauge field. The breaking of the associated $U(1)$ in such a theory is a well-studied phenomenon, which we will not pursue here.

5.2.5 The role of instantons

There is another way to think about the breaking of the $U(1)$ symmetry and θ-dependence in this theory. If one considers the Euclidean functional integral, it is

natural to look for stationary points of the integration, i.e. for classical solutions of the Euclidean equations of motion. In order that they be potentially important, it is necessary that these solutions have finite action, which means that they must be localized in Euclidean space and time. For this reason, such solutions were dubbed "instantons" by 't Hooft. Such solutions are not difficult to find in the CP^N model; we will describe them below. These solutions carry non-zero values of the topological charge,

$$\frac{1}{2\pi}\int d^2x \epsilon_{\mu\nu}F_{\mu\nu} = n, \tag{5.54}$$

and have an action $2\pi n$. If we write $z_i = z_{i\mathrm{cl}} + \delta z_i$, the functional integral, in the presence of a θ term, has the form:

$$Z_{\mathrm{inst}} = e^{\frac{-2\pi n}{g^2}} e^{in\theta}\int [d\delta z_j] e^{-\delta z_i \frac{\delta^2 S}{\delta z_i \delta z_j}\delta z_j} + \cdots . \tag{5.55}$$

It is easy to construct the instanton solution in the case of CP^1. Rather than write the theory in terms of a gauge field, as we have done above, it is convenient to parameterize the theory in terms of a single complex field, Z. One can, for example, define $Z = z_1/z_2$, and let $\bar{Z}$ denote its complex conjugate. Then, with a bit of algebra, one can show that the action for Z which follows from Eq. (5.45) takes the form (it is easiest to work backwards, starting with the equation below and deriving Eq. (5.45)):

$$\mathcal{L} = \frac{\partial_\mu Z \partial_\mu \bar{Z}}{(1+\bar{Z}Z)^2}. \tag{5.56}$$

The function

$$g_{Z\bar{Z}} = \frac{1}{(1+\bar{Z}Z)^2} \tag{5.57}$$

has an interesting significance. There is a well-known mapping of the unit sphere, $x_1^2 + x_2^2 + x_3^2$, onto the complex plane:

$$z = \frac{x_1 + ix_2}{1 - x_3}. \tag{5.58}$$

The inverse is

$$x_1 = \frac{z+z^*}{1+|z|^2} \quad x_2 = \frac{z-z^*}{i(1+|z|^2)} \quad x_3 = \frac{|z|^2-1}{|z|^2+1}. \tag{5.59}$$

The line element on the sphere is mapped in a non-trivial way onto the plane:

$$ds^2 = dx_1^2 + dx_2^2 + dx_3^3 = g_{z\bar{z}}dzd\bar{z} = \frac{1}{(1+\bar{z}z)^z}dzd\bar{z}. \tag{5.60}$$

So the model describes a field constrained to move on a sphere; g is the metric of the sphere. In general, such a model is called a non-linear sigma model. This is an example of a Kahler geometry, a type of geometry which will figure significantly in our discussion of string compactification.

It is straightforward to write down the equations of motion:

$$\partial^2 Z g_{Z\bar{Z}} + \partial_\mu Z \left(\partial_\mu \bar{Z} \frac{\partial g}{\partial \bar{Z}} + \partial_\mu \phi \frac{\partial g}{\partial Z} \right) = 0, \tag{5.61}$$

or

$$\partial_z \partial_{\bar{z}} Z - \frac{2\partial_z Z \partial_{\bar{z}} \bar{Z}}{1 + \bar{Z} Z} = 0. \tag{5.62}$$

Now calling the space-time coordinates $z = x_1 + ix_2$, $z^* = x_1 - ix_2$, we see that if Z is analytic, the equations of motion are satisfied! So a simple solution, which you can check has finite action, is

$$Z(\bar{z}) = \rho \bar{z}. \tag{5.63}$$

In addition to evaluating the action, you can evaluate the topological charge,

$$\frac{1}{2\pi} \int d^2x \epsilon_{\mu\nu} F^{\mu\nu} = 1 \tag{5.64}$$

for this solution. More generally, the topological charge measures the number of times that Z maps the complex plane into the complex plane; $Z = z^n$ has charge n.

We can generalize these solutions. The solution of Eq. (5.63) breaks several symmetries of the action: translation invariance, two-dimensional rotational invariance, and the scale invariance of the classical equations. So we should be able to generate new solutions by translating, rotating and dilating the solution. You can check that

$$Z(z) = \frac{az + b}{cz + d} \tag{5.65}$$

is a solution with action 2π. The parameters $a, \ldots, d$ are called collective coordinates. They correspond to the symmetries of translations, dilations, and rotations, and special conformal transformations (forming the group $SL(2, C)$). In other words, any given finite-action solution breaks the symmetries. In the path integral, the symmetry of Green's functions is recovered when one integrates over the collective coordinates. For translations, this is particularly simple. Integrating over X_0, the instanton position,

$$\langle Z(x) Z(y) \rangle \approx \int d^2x_0 \phi_{\rm cl}(x - x_0) \phi_{\rm cl}(y - x_0) e^{-S_0}. \tag{5.66}$$

(The precise measure is obtained by the Faddeev–Popov method.) Similarly, the integration over the parameter ρ yields a factor

$$\int d\rho \rho^{-1} e^{-\frac{2\pi}{g^2(\rho)}} \dots \tag{5.67}$$

Here the first factor follows on dimensional grounds. The second follows from renormalization-group considerations. It can be found by explicit evaluation of the functional determinant. Note that, because of asymptotic freedom, this means that typical Green functions will be divergent in the infrared.

There are many other features of this instanton one can consider. For example, one can add massless fermions to the model. The resulting theory has a chiral $U(1)$ symmetry, which is anomalous. The instanton gives rise to non-zero Green functions which violate the $U(1)$ symmetry. We will leave investigation of fermions in this model to the exercises, and turn to the theory of interest, which exhibits phenomena parallel to this simple theory.

5.3 Real QCD

The model of the previous section mimics many features of real QCD. Indeed, we will see that much of our discussion can be carried over, almost word for word, to the observed strong interactions. This analogy is helpful, given that in QCD we have no approximation which gives us control over the theory comparable to that which we found in the large-N limit of the CP^N model. As in that theory, we have the following.

- There is a θ parameter, which appears as an integral over the divergence of a non-gauge invariant current.
- There are instantons, which indicate that physical quantities should be θ-dependent. However, instanton effects cannot be considered in a controlled approximation, and there is no clear sense in which θ-dependence can be understood as arising from instantons.
- In QCD, there is also a large-N expansion, but while it produces significant simplification, one cannot solve the theory even in the leading large-N approximation. Instead, understanding of the underlying symmetries, and experimental information about chiral symmetry breaking, provides critical information about the behavior of the strongly coupled theory, and allows computations of the physical effects of θ.

5.3.1 The theory and its symmetries

In order to understand the effects of θ, it is sufficient to focus on only the light quark sector of QCD. For simplicity in writing some of the formulas, we will consider two light quarks; it is not difficult to generalize the resulting analysis to the case of

three. It is believed that the masses of the u and d quarks are of order 5 MeV and 10 MeV, respectively, much smaller than the scale of QCD. So we first consider an idealization of the theory in which these masses are set to zero. In this limit, the theory has a symmetry $SU(2)_{\rm L} \times SU(2)_{\rm R}$. Calling

$$q = \begin{pmatrix} u \\ d \end{pmatrix} \quad \bar{q} = \begin{pmatrix} \bar{u} \\ \bar{d} \end{pmatrix}, \tag{5.68}$$

the two $SU(2)$ symmetries act separately on q and $\bar{q}$ (thought of as left-handed fermions),

$$q^T \to q^T U_{\rm L} \quad \bar{q} \to U_{\rm R} \bar{q}. \tag{5.69}$$

This symmetry is spontaneously broken. The order parameter for the symmetry breaking is believed to be an expectation value for the quark bilinear:

$$\mathcal{M} = \bar{q} q. \tag{5.70}$$

Under the original symmetry,

$$\mathcal{M} \to U_{\rm R} \mathcal{M} U_{\rm L}. \tag{5.71}$$

The expectation value (condensate) of $\mathcal{M}$ is

$$\mathcal{M} = c \Lambda_{\rm qcd}^3 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{5.72}$$

This breaks some of the original symmetry, but preserves the symmetry $U_{\rm L} = U_{\rm R}$. This symmetry is just the $SU(2)$ of isospin. The Goldstone bosons associated with the three broken symmetry generators must transform in a representation of the unbroken symmetry: these are just the pions, which are a vector of isospin. One can think of the Goldstone bosons as being associated with a slow variation of the expectation value in space, so we can introduce a composite operator

$$\mathcal{M} = \bar{q} q = M_0 e^{i \frac{\pi_a(x) \tau_a}{f_\pi}} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{5.73}$$

The quark mass term in the Lagrangian is then (for simplicity taking $m_u = m_d = m_q$)

$$m_q \mathcal{M}. \tag{5.74}$$

Replacing $\mathcal{M}$ by its expression (5.73) gives a potential for the pion fields. Expanding $\mathcal{M}$ in powers of π / f_π, the minimum of the potential occurs for $\pi_a = 0$. Expanding to second order, one has

$$m_\pi^2 f_\pi^2 = m_q M_0. \tag{5.75}$$

We have been a bit cavalier about the symmetries. The theory also has two $U(1)$ symmetries:

$$q \to e^{i\alpha} q \quad \bar{q} \to e^{i\alpha} \bar{q} \tag{5.76}$$

$$q \to e^{i\alpha} q \quad \bar{q} \to e^{-i\alpha} \bar{q} \tag{5.77}$$

The first of these is baryon number and it is not chiral (and is not broken by the condensate). The second is the axial $U(1)_5$; it is also broken by the condensate. So, in addition to the pions, there should be another approximate Goldstone boson. But there is no good candidate among the known hadrons. The η, has the right quantum numbers, but, as we will see below, the η is too heavy to be interpreted in this way. The absence of this fourth (or, in the case of three light quarks, ninth) Goldstone boson is called the $U(1)$ problem.

The $U(1)_5$ symmetry suffers from an anomaly, however, and we might hope that this has something to do with the absence of a corresponding Goldstone boson. The anomaly is given by

$$\partial_\mu j_5^\mu = \frac{1}{16\pi^2} F\tilde{F}. \tag{5.78}$$

Again, we can write the right-hand side as a total divergence,

$$F\tilde{F} = \partial_\mu K^\mu, \tag{5.79}$$

where

$$K_\mu = \epsilon_{\mu\nu\rho\sigma} \left(A_\nu^a F_{\rho\sigma}^a - \frac{2}{3} f^{abc} A_\nu^a A_\rho^b A_\sigma^c \right). \tag{5.80}$$

This accounts for the fact that in perturbation a theory the axial $U(1)$ is conserved. Non-perturbatively, as we will now show, there are important configurations in the functional integral for which the right-hand side does not vanish rapidly at infinity.

5.3.2 Instantons in QCD

In the Euclidean functional integral

$$Z = \int [dA][dq][d\bar{q}] e^{-S} \tag{5.81}$$

it is natural to look for stationary points of the effective action, i.e. finite action, classical solutions of the theory in imaginary time. The Yang–Mills equations are complicated, non-linear equations, but it turns out that, much as in the CP^N model, the instanton solutions can be found rather easily. The following tricks simplify the construction, and turn out to yield the general solution. First, note that the

Yang–Mills action satisfies an inequality, the Bogomolnyi bound:

$$\int (F \pm \tilde{F})^2 = \int (F^2 + \tilde{F}^2 \pm 2F\tilde{F}) = \int (2F^2 + 2F\tilde{F}) \geq 0. \tag{5.82}$$

So the action is bounded by $\int F\tilde{F}$, with the bound being saturated when

$$F = \pm\tilde{F} \tag{5.83}$$

i.e. if the gauge field is (anti-) self-dual.[1] This equation is a first-order equation, and it is easy to solve if one first restricts to an $SU(2)$ subgroup of the full gauge group. One makes the ansatz that the solution should be invariant under a combination of ordinary rotations and global $SU(2)$ gauge transformations:

$$A_\mu = f(r^2) + h(r^2)\vec{x} \cdot \vec{\tau} \tag{5.84}$$

where we are using the matrix notation for the gauge fields. One can actually make a better guess: define the gauge transformation

$$g(x) = \frac{x_4 + i\vec{x} \cdot \vec{\tau}}{r} \tag{5.85}$$

and take

$$A_\mu = f(r^2) g \partial_\mu g^{-1}. \tag{5.86}$$

Then plugging in the Yang–Mills equations yields:

$$f = \frac{-ir^2}{r^2 + \rho^2} \tag{5.87}$$

where ρ is an arbitrary quantity with dimensions of length. The choice of origin here is also arbitrary; this can be remedied by simply replacing $x \to x - x_0$ everywhere in these expressions, where x_0 represents the location of the instanton.

From this solution, it is clear why $\int \partial_\mu K^\mu$ does not vanish for the solution: while A is a pure gauge at infinity, it falls only as $1/r$. Indeed, since $F = \tilde{F}$, for this solution

$$\int F^2 = \int \tilde{F}^2 = 32\pi^2. \tag{5.88}$$

This result can also be understood topologically. Note that g defines a mapping from the "sphere at infinity" into the gauge group. It is straightforward to show that

$$\frac{1}{32\pi^2} \int d^4x F\tilde{F} \tag{5.89}$$

[1] This is not an accident, nor was the analyticity condition in the CP^N case. In both cases, we can add fermions so that the model is supersymmetric. Then one can show that if some of the supersymmetry generators, Q_α, annihilate a field configuration, then the configuration is a solution. This is a first-order condition; in the Yang–Mills case, it implies self-duality, and in the CP^N case it requires analyticity.

counts the number of times g maps the sphere at infinity into the group (one for this specific example; n more generally). In the exercises and suggested reading, features of the instanton are explored in more detail.

So we have exhibited potentially important contributions to the path integral which violate the $U(1)$ symmetry. How does this violation of the symmetry show up? Let's consider the path integral more carefully. Having found a classical solution, we want to integrate about small fluctuations about it. Including the θ term, these have the form

$$\langle \bar{u}u\bar{d}d\rangle = e^{-\frac{8\pi^2}{g^2}} e^{i\theta} \int [d\delta A][dq][d\bar{q}] \exp\left(-\frac{\delta^2 S}{\delta A^2}\delta A^2 - S_{q,\bar{q}}\right) \bar{u}u\bar{d}d. \quad (5.90)$$

Now S contains an explicit factor of $1/g^2$. As a result, the fluctuations are formally suppressed by g^2 relative to the leading contribution. The one-loop functional integral yields a product of determinants for the fermions, and of inverse square root determinants for the bosons.

Consider the integral over the fermions. It is straightforward, if challenging, to evaluate the determinants. But if the quark masses are zero, the fermion functional integrals are zero, because there is a zero mode for each of the fermions, i.e. for both q and $\bar{q}$ there is a normalizable solution of the equation:

$$\not{D}u = 0 \quad \not{D}\bar{u} = 0 \quad (5.91)$$

and similarly for d and $\bar{d}$. It is straightforward to construct these solutions:

$$u = \frac{\rho}{(\rho^2 + (x - x_0)^2)^{3/2}}\zeta, \quad (5.92)$$

where ζ is a constant spinor, and similarly for $\bar{u}$, etc.

Let's understand this a bit more precisely. Euclidean path integrals are conceptually simple. Consider some classical solution, $\Phi_{\rm cl}(x)$ (here Φ denotes collectively the various bosonic fields; we will treat, for now, the fermions as vanishing in the classical solutions). In the path integral, at small coupling, we are interested in small fluctuations about the classical solution,

$$\Phi = \Phi_{\rm cl} + \delta\Phi. \quad (5.93)$$

Because the action is stationary at the classical solution,

$$S = S_{\rm cl} + \int d^4x \delta\Phi \frac{\partial^2 \mathcal{L}}{\partial \Phi^2}\delta\Phi + \cdots. \quad (5.94)$$

The second derivative here is a shorthand for a second-order differential operator, which we will simply denote by S'', and refer to as the *quadratic fluctuation operator*. We can expand $\delta\Phi$ in (normalizable) eigenfunctions of this operator Φ_n with eigenvalues λ_n, $\Phi = c_n\Phi_n$. The result of the functional integral is then $\prod \lambda_n^{-1/2}$.

This is the leading correction to the classical limit. Higher-order corrections are suppressed by powers of g^2. This is most easily seen by working in the scaling where the action has a $1/g^2$ out front. Then one can derive the perturbation theory from the path integral in the usual way; the main difference from the usual treatment with zero background fields is that the propagators are more complicated. The propagators for various fields in the instanton background are in fact known in closed form.

The form of the differential operator is familiar from our calculation of the β-function in the background field method (in the background field gauge). For the gauge bosons, in a suitable ("background field") gauge, it is

$$S'' = \mathcal{D}^2 + \mathcal{J}_{\mu\nu} F^{\mu\nu}. \tag{5.95}$$

Here $\mathcal{D}$ is just the covariant derivative with the vector potential corresponding to the classical solution (instanton), and similarly for the field strength; $\mathcal{J}_{\mu\nu}$ is the generator of Lorentz transformations in the vector representation. The eigenvalue problem was completely solved by 't Hooft.

Both the bosonic and fermionic quadratic fluctuation operators have zero eigenvalues. For the bosons, these potentially give infinite contributions to the functional integral, and they must be treated separately. The difficulty is that among the variations of the fields are symmetry transformations: changes in the location of the instanton (translations), rotations of the instanton, and scale transformations. Consider translations. For every solution, there is an infinite set of solutions obtained by shifting the origin (varying x_0). Instead of integrating over a coefficient, c_0, we integrate over the *collective coordinate* x_0 (one must also include a suitable Jacobian factor). The effect of this is to restore translational invariance in Green's functions. We will see this explicitly shortly. Similarly, the instanton breaks the rotational invariance of the theory. Correspondingly, we can find a three-parameter set of solutions and zero modes. Integrating over these rotational collective coordinates restores rotational invariance. (The instanton also breaks a global gauge symmetry, but a combination of rotations and gauge transformations is preserved.)

Finally, the classical theory is scale invariant; this is the origin of the parameter ρ in the solution. Again, one must treat ρ as a collective coordinate, and integrate over ρ. There is a power of ρ arising from the Jacobian, which can be determined on dimensional grounds. For the Green function, Eq. (5.90), for example, which has dimension six, we have (if all of the fields are evaluated at the same point),

$$\int d\rho \rho^{-7}. \tag{5.96}$$

However, there is additional ρ-dependence because the quantum theory violates the scale symmetry. This can be understood by replacing $g^2 \to g^2(\rho)$ in the functional

integral, and using

$$e^{-8\pi^2 g^2(\rho)} \approx (\rho M)^{b_0} \tag{5.97}$$

for small ρ. For three-flavor QCD, for example, $b_0 = 9$, and the ρ integral diverges for large ρ. This is just the statement that the integral is dominated by the infrared, where the QCD coupling becomes strong.

Fermion functional integrals introduce a new feature. In four-component language, it is necessary to treat q and $\bar{q}$ as independent fields. This rule gives the functional integral as a determinant, rather than as, say, the square root of a determinant. (In two-component language, this corresponds to treating q and q^* as independent fields.) So at one-loop order, we need to study:

$$\not{D} q_n = \lambda_n q_n \qquad \not{D} \bar{q}_n = \lambda_n \bar{q}_n \tag{5.98}$$

For non-zero λ_n there is a pairing of solutions with opposite eigenvalues of γ_5. In four-component notation, one can see this from:

$$\not{D} q_n = \lambda_n q_n \;\rightarrow\; \not{D}\gamma_5 q_n = -\lambda_n \gamma_5 q_n. \tag{5.99}$$

Zero eigenvalues, however, are special. There is no corresponding pairing. This has implications for the fermion functional integral. Writing

$$q(x) = \sum a_n q_n(x), \tag{5.100}$$

$$S = \sum \lambda_n a_n^* a_n. \tag{5.101}$$

Then

$$\int [dq][d\bar{q}] e^{-S} = \prod_{n=0}^{\infty} da_n da_n^* e^{-\sum_{n\neq 0} \lambda_n a_n^* a_n}. \tag{5.102}$$

Because the zero modes do not contribute to the action, many Green functions vanish. For example, $\langle 1 \rangle = 0$. In order to obtain a non-vanishing result, we need enough insertions of q to "soak up" all of the zero modes.

We have seen that, in the instanton background, there are normalizable fermion zero modes, one for each left-handed field. This means that in order for the path integral to be non-vanishing, we need to include insertions of enough qs and $\bar{q}$s to soak up all of the zero modes. In other words, in two-flavor QCD, non-vanishing Green functions have the form

$$\langle \bar{u} u \bar{d} d \rangle \tag{5.103}$$

and violate the symmetry. Note that the symmetry violation is just as predicted from the anomaly equation:

$$\Delta Q_5 = \frac{2}{16\pi^2} \int d^4x F\tilde{F} = 4. \tag{5.104}$$

This is a particular example of an important mathematical theorem known as the Atiyah–Singer index theorem.

We can put all of this together to evaluate a Green function which violates the classical $U(1)$ symmetry of the massless theory, $\langle \bar{u}(x)u(x)\bar{d}(x)d(x)\rangle$. Taking the gauge group to be $SU(2)$, there is one zero mode for each of u, $\bar{u}$, d and $\bar{d}$. The fields in this expectation value can soak up all of these zero modes. The effect of the integration over x_0 is to give a result independent of x, since the zero modes are functions only of $x - x_0$. The integration over the rotational zero modes gives a non-zero result only if the Lorentz indices are contracted in a rotationally invariant manner (the same applies to the gauge indices). The integration over the instanton scale size – the conformal collective coordinate – is more problematic, exhibiting precisely the infrared divergence of Eq. (5.96).

So we have provided some evidence that the $U(1)$ problem is solved in QCD, but no reliable calculation. What about θ-dependence? Let us ask first about θ-dependence of the vacuum energy. In order to get a non-zero result, we need to allow that the quarks are massive. Treating the mass as a perturbation, we obtain a result of the form:

$$E(\theta) = C\Lambda_{\rm QCD}^9 m_u m_d \cos(\theta) \int d\rho \rho^{-3} \rho^9. \tag{5.105}$$

So, as in the CP^N model, we have evidence for θ-dependence, but cannot do a reliable calculation. That we cannot do a calculation should not be a surprise. There is no small parameter in QCD to use as an expansion parameter. Fortunately, we can use other facts which we know about the strong interactions to get a better handle on both the $U(1)$ problem and the question of θ-dependence.

Before continuing, however, let us consider the weak interactions. Here there is a small parameter, and there are no infrared difficulties, so we might expect instanton effects to be small. The analog of the $U(1)_5$ symmetry in this case is baryon number. Baryon number has an anomaly in the standard model, since all of the quark doublets have the same sign of the baryon number. 't Hooft showed that one could actually use instantons, in this case, to compute the violation of baryon number. Technically, there are no finite-action Euclidean solutions in this theory; this follows, as we will see in a moment, from a simple scaling argument. However, 't Hooft realized that one can construct important configurations of non-zero topological charge by starting with the instantons of the pure gauge theory and perturbing them. For the Higgs boson, one solves the equation

$$D^2\phi = V'(\phi). \tag{5.106}$$

For a light boson, one can neglect the right-hand side. Then this equation is solved by:

$$\phi(x) = i\bar{\sigma}^{\mu} x^{\mu} \left(\frac{1}{x^2 + \rho^2} \right)^{\frac{1}{2}} \langle \phi \rangle. \tag{5.107}$$

Note that at large x, this has the form $g(x)\langle\phi\rangle$. As a result, the action of the configuration is finite. One finds a correction to the action

$$\delta S = \frac{1}{g^2} v^2 \rho^2. \tag{5.108}$$

Including this in the exponential damps the ρ integral at large ρ, and leads to a convergent result.

Including now the fermions, there is a zero mode for each $SU(2)$ doublet. So one obtains a non-zero expectation value for correlation functions of the form $\langle QQQLLL \rangle$, where the color and $SU(2)$ indices are contracted in a gauge invariant way, and the flavors for the Qs and Ls are all different. The coefficient has the form:

$$\mathcal{A}_{\rm bv} = C\, e^{-\frac{2\pi}{\alpha_{\rm w}}}. \tag{5.109}$$

From this, one can see that baryon number violation occurs in the Standard Model, but at an incredibly small rate. One can also calculate a term in the effective action, involving three quarks and three leptons, with a similar coefficient, by studying Green functions in which all of the fields are widely separated. We will encounter this sort of computation later, when we discuss instantons in supersymmetric theories.

5.3.3 Physical interpretation of the instanton solution

We have derived dramatic physical effects from the instanton solution by direct calculation, but we have not provided a physical picture of the phenomena the instanton describes. Already in quantum mechanics, imaginary time solutions of the classical equations of motion are familiar in the WKB analysis of tunneling, and the Yang–Mills instanton (and the CP^N instanton) also describe tunneling phenomena. In this section, we will confine our attention to pure gauge theories. The generalization to theories with fermions and/or scalars is straightforward and interesting.

To understand the instanton as tunneling, it is helpful to work in a non-covariant gauge, in which there is a Hamiltonian description. The gauge $A_0 = 0$ is particularly useful. In this gauge, the canonical coordinates are the A_is, and their conjugate momenta are E_is (with a minus sign). This is too many degrees of freedom if all are treated as independent. The resolution lies in the need to enforce

Gauss's law, which is now to be viewed as an operator constraint on states. For example, in a $U(1)$ theory,

$$G(\vec{x})|\Psi\rangle = (\vec{\nabla}\cdot\vec{E} - \rho)|\Psi\rangle = 0. \tag{5.110}$$

The left-hand side is almost the generator of gauge transformations. On the gauge fields, for example,

$$\left[\int d^3x\omega(\vec{x})G(\vec{x}),\, A_i(\vec{y})\right] = -\int d^3x\,\partial_j\omega(\vec{x})[E(\vec{x})_j,\, A(\vec{y})_i] = \partial_i\omega(\vec{y}). \tag{5.111}$$

In the second step we integrated by parts and dropped a possible surface term. This requires that $\omega \to 0$ fast enough at infinity. Such gauge transformations are called "small" gauge transformations. We have learned that in $A_0 = 0$ gauge, states must be invariant under time-independent, small gauge transformations.

In electrodynamics, this is not particularly interesting. But the same manipulations hold in non-Abelian theories, and in this case there are interesting large gauge transformations. An example is

$$g(\vec{x}) = \exp\left(i\pi\frac{\vec{x}\cdot\vec{\sigma}}{\sqrt{\vec{x}^2 + a^2}}\right). \tag{5.112}$$

We can also consider powers of g, g^n. We can think of g as mapping the three-dimensional space into the group $SU(2)$. The number of times that the mapping wraps around the gauge group is known as the winding number, and it can be written as:

$$n = \frac{1}{24\pi^2}\int d^3x\epsilon_{ijk}\,\mathrm{Tr}[\partial_i g\partial_j g\partial_k g]. \tag{5.113}$$

However, g_n is not unique; we can multiply by any small gauge transformation without changing n. The zero energy states consist of $A_i = ig^{-n}\partial_i g^n$, averaged over the small gauge transformations so as to make them invariant.

With just a little algebra, one can show that $n = \int d^3x\,K_0$, where K^μ is the topological current we encountered in Eq. (5.80). So an instanton, in $A_0 = 0$ gauge, corresponds to a tunneling between states of different n. More precisely, there is a non-zero matrix element of the Hamiltonian between states of different n,

$$\langle n|H|n \pm 1\rangle = \epsilon. \tag{5.114}$$

This is analogous to the situation in crystals, and the energy eigenstates are similar to Bloch waves:

$$|\theta\rangle = \sum_n e^{in\theta}|n\rangle, \tag{5.115}$$

with energy $\epsilon \cos(\theta)$. This θ is precisely the same θ which entered as a parameter in the Lagrangian.

5.3.4 QCD and the $U(1)$ problem

In real QCD, we have seen that, on the one hand, instanton configurations violate the axial $U(1)$ symmetry. On the other hand, the explicit calculations are infrared-divergent, so we cannot make a reliable calculation of the Goldstone mass. This is not a surprise; there is no small parameter which would justify the use of a semi-classical approximation. Still, the instanton analysis we have described makes clear that there is no reason to expect that there is a light Goldstone boson. Actually, while perturbative and semiclassical (instanton) techniques have no reason to give reliable results, there are two approximation methods techniques which are available. The first is large N, where one now allows the N of $SU(N)$ to be large, with g^2N fixed. In contrast to the case of CP^N, this does not permit enough simplification to permit explicit computations, but it does allow one to make qualitative statements about the theory. Witten has pointed out a way in which one can relate the mass of the η (or η' if one is thinking in terms of $SU(3) \times SU(3)$ current algebra) to quantities in a theory without quarks. The anomaly is an effect suppressed by a power of N, in the large N limit. This is because the loop diagram contains a factor of g^2 but not of N. So, in large N, it can be treated as a perturbation, and the η is almost massless. The $\partial_\mu j_5^\mu$ is like a creation operator for η, so (just like $\partial_\mu j_5^{\mu\ 3}$ is a creation operator for the π meson), so one can compute the mass if one knows the correlation function, at zero momentum,

$$\left\langle \partial_\mu j_5^\mu(x) \partial_\mu j_5^\mu(y) \right\rangle \propto \frac{1}{N^2} \langle F(x)\tilde{F}(x)F(y)\tilde{F}(y)\rangle. \tag{5.116}$$

To leading order in the $1/N$ expansion, the $F\tilde{F}$ correlation function can be computed in the theory without quarks. Witten argued that while this vanishes order by order in perturbation theory, there is no reason that this correlation function need vanish in the full theory. Attempts have been made to compute this quantity both in lattice gauge theory and using the AdS–CFT correspondence recently discovered in string theory. Both methods give promising results.

So the $U(1)$ problem should be viewed as solved, in the sense that absent any argument to the contrary, there is no reason to think that there should be an extra Goldstone boson in QCD.

The second approximation scheme which gives some control of QCD is known as chiral perturbation theory. The masses of the u, d and s quarks are small compared to the QCD scale, and the mass terms for these quarks in the Lagrangian can be treated as perturbations. This will figure in our discussion in the next section.

5.4 The strong CP problem

5.4.1 The θ-dependence of the vacuum energy

The fact that the anomaly resolves the $U(1)$ problem in QCD raises another issue. Given that $\int d^4x F\tilde{F}$ has physical effects, a theta term in the action has physical effects as well. Since this term is CP odd, this means that there is the potential for strong CP-violating effects. These effects should vanish in the limit of zero quark masses, since in this case, by a field redefinition, we can remove θ from the Lagrangian. In the presence of quark masses, the θ-dependence of many quantities can be computed. Consider, for example, the vacuum energy. In QCD, the quark mass term in the Lagrangian has the form:

$$\mathcal{L}_{\rm m} = m_u \bar{u}u + m_d \bar{d}d + \text{h.c.} \tag{5.117}$$

Were it not for the anomaly, we could, by redefining the quark fields, take m_u and m_d to be real. Instead, we can define these fields so that there is no $\theta F\tilde{F}$ term in the action, but there is a phase in m_u and m_d. Clearly, we have some freedom in making this choice. In the case that m_u and m_d are equal, it is natural to choose these phases to be the same. We will explain shortly how one proceeds when the masses are different (as they are in nature). So

$$\mathcal{L}_{\rm m} = (m_u \bar{u}u + m_d \bar{d}d)e^{i\theta} + \text{h.c.} \tag{5.118}$$

Now we want to treat this term as a perturbation. At first order, it makes a contribution to the ground-state energy proportional to its expectation value. We have already argued that the quark bilinears have non-zero vacuum expectation values, so

$$E(\theta) = (m_u + m_d)\cos(\theta)\langle \bar{q}q \rangle. \tag{5.119}$$

While, without a difficult non-perturbative calculation, we can't calculate the separate quantities on the right-hand side of this expression, we can, using current algebra, relate them to measured quantities. We have seen (Appendix B) that

$$m_{\pi^2} f_{\pi^2} = \text{Tr}\, M_q \langle \Sigma \rangle = (m_u + m_d)\langle \bar{q}q \rangle. \tag{5.120}$$

Replacing the quark mass terms in the Lagrangian by their expectation values, we can immediately read off the energy of the vacuum as a function of θ:

$$E(\theta) = m_\pi^2 f_\pi^2 \cos(\theta). \tag{5.121}$$

This expression can readily be generalized to the case of three light quarks by similar methods. So we see that there is real physics in θ, even if we don't understand how to do an instanton calculation. In the next section, we will calculate a more interesting quantity: the neutron electric dipole moment as a function of θ.

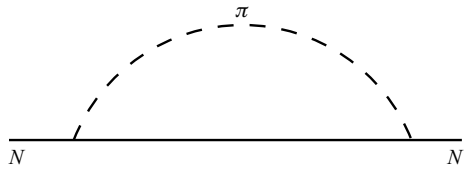

Fig. 5.2. Diagram in which CP-violating coupling of the pion contributes to $d_{\rm n}$.

5.4.2 The neutron electric dipole moment

The most interesting physical quantities to study in connection with CP violation are electric dipole moments, particularly that of the neutron, $d_{\rm n}$. If CP were badly violated in the strong interactions, one might expect $d_{\rm n} \approx e\,\text{fm} \approx 10^{-14}$ cm. But the experimental limits on the dipole moment are striking,

$$d_{\rm n} < 10^{-25}\, e\ \text{cm}. \tag{5.122}$$

Using current algebra, the leading contribution to the neutron electric dipole moment due to θ can be calculated, and one obtains a limit $\theta < 10^{-9}$. Here we outline the main steps in the calculation; I urge you to work out the details following the reference in the suggested reading. We will simplify the analysis by working in an exact $SU(2)$-symmetric limit, i.e. by taking $m_u = m_d = m$. We again treat the Lagrangian of Eq. (5.118) as a perturbation. We can understand how this term depends on the π fields by making an axial $SU(2)$ transformation on the quark fields. In other words, a background π field can be thought of as a small chiral transformation on the vacuum. Then, for example, for the τ_3 direction, $q \to (1 + i\pi_3\tau_3)q$ (the π field parameterizes the transformation), so the action becomes:

$$\frac{m}{f_\pi}\pi_3(\bar{q}\gamma_5 q + \theta\bar{q}q). \tag{5.123}$$

The second term is a CP-violating coupling of the mesons to the pions.

This coupling is difficult to measure directly, but this coupling gives rise, in a calculable fashion, to a neutron electric dipole moment. Consider the graph of Fig. 5.2. This graph generates a neutron electric dipole moment, if we take one coupling to be the standard pion–nucleon coupling, and the second the coupling we have computed above. The resulting Feynman graph is infrared divergent; we cut this off at m_π, while cutting off the integral in the ultraviolet at the QCD scale. The low-energy calculation is reliable in the limit that m_π is small, so that $\ln(m_\pi/\Lambda_{\rm QCD})$ is large compared to one. The result is:

$$d_{\rm n} = g_{\pi NN}\frac{-\theta m_u m_d}{f_\pi(m_u + m_d)}\langle N_f|\bar{q}\tau^a q|N_i\rangle \ln(M_N/m_\pi)\frac{1}{4\pi^2}M_N. \tag{5.124}$$

The matrix element can be estimated using the $SU(3)$ symmetry of Gell-Mann and Ne'eman, yielding $d_n = 5.2 \times 10^{-16}\theta$ cm. The experimental bound gives $\theta < 10^{-9}$–10^{-10}. Understanding why CP violation is so small in the strong interactions is the "strong CP problem."

5.5 Possible solutions of the strong CP problem

What should our attitude towards this problem be? We might argue that, after all, some Yukawa couplings are as small as 10^{-5}, so why is 10^{-9} so bad? On the other hand, we suspect that the smallness of the Yukawa couplings is related to approximate symmetries, and that these Yukawa couplings are telling us something. Perhaps there is some explanation of the smallness of θ, and perhaps this is a clue to new physics. In this section we review some of the solutions which have been proposed to understand the smallness of θ.

5.5.1 When $m_u = 0$

Suppose that the mass of the up quark were zero. In this case, by a field redefinition of the up quark,

$$u \to e^{-i\theta} u, \tag{5.125}$$

one could make the θ term vanish as a consequence of the anomaly. This is a simple enough explanation, but there are two issues. First, why? We could imagine that this is the result of a symmetry, but this symmetry cannot be a real symmetry of the underlying theory, since it is violated by QCD (through the anomaly). We will see later in this book that discrete symmetries, with anomalies of the kind required to understand a vanishing u quark mass, *do* frequently arise in string theory. So perhaps this sort of explanation is plausible.

There is a much more practical issue which arises from the phenomenology of QCD itself. Standard studies of QCD current algebra give the non-zero value of m_u which we quoted earlier, around 5 MeV. This, by itself, is not necessarily a serious objection. The question is: at what momentum transfer is the mass evaluated? Suppose, for example, that this is the u quark mass at 300 MeV, while the mass of the u quark at, say, 10 GeV is essentially zero. Integrating out physics between the scale 10 GeV and 0.3 GeV will induce a u quark mass in the effective Lagrangian. Instantons will generate such a term. It will be necessary to soak up the d and s quark zero modes with the d and s quark mass terms, so the u quark mass will be proportional to $m_d m_s$. We can't actually do a reliable calculation, but we would expect:

$$m_u = \frac{m_d m_s}{300 \text{ MeV}}, \tag{5.126}$$

which is not much different than the usual current algebra number. At this point, however, numerical lattice calculations are good enough to address this issue, and a massless u quark at several GeV appears to be incompatible with the data.

5.5.2 Spontaneous CP violation

Suppose that the underlying theory respects CP, and the observed CP violation is spontaneous. Because θ is CP odd, the underlying theory has $\theta = 0$. One might hope that this feature would be preserved when the symmetry is spontaneously broken.

There are a number of ways that θ might be generated in the low-energy theory. First, suppose that CP is broken by the expectation value of a complex field, Φ. There might well be direct couplings, such as

$$\frac{1}{16\pi^2}\mathrm{Im}\Phi F\tilde{F}. \tag{5.127}$$

Note that Φ might also couple to fermions, giving them a large mass through its expectation value. When these fermions are integrated out, this would also generate an effective θ. This is likely simply given the anomalous field redefinitions which may be required to make the masses of these fields real. It is possible to construct models where, at least at tree level, the heavy fermion mass matrices are real. Quantum corrections, and the couplings mentioned above, pose additional challenges. It is an open question whether such an explanation of the smallness of θ is plausible, and has testable consequences.

5.5.3 The axion

Perhaps the most compelling explanation of the smallness of θ involves a hypothetical particle called the axion. We present here a slightly updated version of the original idea of Peccei and Quinn.

Consider the vacuum energy as a function of θ (Eq. (5.119)). This energy has a minimum at $\theta = 0$, i.e. at the CP conserving point. As Weinberg noted long ago, this is almost automatic: points of higher symmetry are necessarily stationary points. As it stands, this observation is not particularly useful, since θ is a parameter, not a dynamical variable. But suppose that one has a field, a, with coupling to QCD:

$$\mathcal{L}_{\mathrm{axion}} = (\partial_\mu a)^2 + \frac{(a/f_\mathrm{a} + \theta)}{32\pi^2}F\tilde{F}, \tag{5.128}$$

where f_a is known as the axion decay constant. Suppose, in addition, that the rest of the theory possesses a symmetry, called the Peccei–Quinn symmetry,

$$a \to a + \alpha \tag{5.129}$$

for constant α. Then, by a shift in a, one can eliminate θ. What we previously called the vacuum energy as a function of θ, $E(\theta)$, is now $V(a/f_a)$, the potential energy of the axion. It has a minimum at $\theta = 0$. The strong CP problem is solved.

One can estimate the axion mass by simply examining $E(\theta)$:

$$m_a^2 \approx \frac{m_\pi^2 f_{\pi^2}}{f_a^2}. \tag{5.130}$$

If $f_a \sim$ TeV, this yields a mass of order keV. If $f_a \sim 10^{16}$ GeV, this gives a mass of order 10^{-9} eV.

There are several questions one can raise about this proposal.

- Should the axion already have been observed? The couplings of the axion to matter can be worked out in a given model in a straightforward way, using the methods of current algebra (in particular of non-linear Lagrangians). All of the couplings of the axion are suppressed by powers of f_a. This is characteristic of a Goldstone boson. At zero momentum, a change in the field is like a symmetry transformation, so, before including the QCD effects which explicitly break the symmetry, axion couplings are suppressed by powers of momentum over f_a; QCD effects are suppressed by Λ_{QCD}/f_a. So if f_a is large enough, the axion is difficult to see. The strongest limit turns out to come from red giant stars. The production of axions is "semiweak," i.e. it only is suppressed by one power of f_a, rather than two powers of m_W; as a result, axion emission is competitive with neutrino emission until $f_a > 10^{10}$ GeV or so.
- As we will describe in more detail in the chapters on cosmology, the axion can be copiously produced in the early universe. As a result, there is an *upper bound* on the axion decay constant, of about 10^{11} GeV. If this bound is saturated, the axion constitutes the dark matter. We will discuss this bound in detail in the chapter on particle astrophysics.
- Can one search for the axion experimentally? Typically, the axion couples not only to the $F\tilde{F}$ of QCD, but also to the same object in QED. This means that in a strong magnetic field, an axion can convert to a photon. Precisely this effect is being searched for by groups at Livermore (the collaboration contains members from MIT and the University of Florida as well) and Kyoto. The basic idea is to suppose that the dark matter in the halo of our galaxy consists principally of axions. Using a (superconducting) resonant cavity with a high Q value in a large magnetic field, one searches for the conversion of these axions into excitations of the cavity, owing to the coupling of the axion to the electromagnetic $F\tilde{F} = \vec{E} \cdot \vec{B}$. The experiments have already reached a level where they set interesting limits; the next generation of experiments will cut a significant swath in the presently allowed parameter space.
- The coupling of the axion to $F\tilde{F}$ violates the shift symmetry; this is why the axion can develop a potential. But this seems rather paradoxical: one postulates a symmetry, preserved to some high degree of approximation, but which is not a symmetry; it is at least broken by tiny QCD effects. Is this reasonable? To understand the nature of the problem, consider one of the ways an axion can arise. In some approximation, we

can suppose we have a global symmetry under which a scalar field, ϕ, transforms as $\phi \to e^{i\alpha}\phi$. Suppose, further, that ϕ has an expectation value. This could arise due to a potential, $V(\phi) = -\mu^2|\phi|^2 + \lambda|\phi|^4$. Associated with the symmetry breaking would be a (pseudo)-Goldstone boson, a. We can parameterize ϕ as:

$$\phi = f_a e^{ia/f_a} \quad |\langle\phi\rangle| = f_a. \tag{5.131}$$

If this field couples to fermions, they gain mass from its expectation value. At one loop, the same diagrams we discussed in our anomaly analysis generate a coupling $aF\tilde{F}$, from integrating out the fermions. This calculation is identical to the corresponding calculation for pions we discussed earlier. But we usually assume that global symmetries in nature are accidents. For example, baryon number is conserved in the standard model simply because there are no gauge-invariant, renormalizable operators which violate the symmetry. We believe it is violated by higher dimension terms. The global symmetry we postulate here is presumably an accident of the same sort. But for the axion, the symmetry must be *extremely* good. For example, suppose one has a symmetry breaking operator

$$\frac{\phi^{n+4}}{M_p^n}. \tag{5.132}$$

Such a term gives a linear contribution to the axion potential of order f_a^{n+3}/M_p^n. If $f_a \sim 10^{11}$, this swamps the would-be QCD contribution ($m_\pi^2 f_\pi^2/f_a$) unless $n > 12$!

This last objection finds an answer in string theory. In this theory, there are axions, with just the right properties, i.e. there are symmetries in the theory which are *exact* in perturbation theory, but which are broken by exponentially small non-perturbative effects. The most natural value for f_a would appear to be of order $M_{GUT} - M_p$. Whether this can be made compatible with cosmology, or whether one can obtain a lower scale, is an open question.

Suggested reading

There are a number of excellent books and reviews on anomalies, as well as good treatments in quantum field theory textbooks. The texts of Peskin and Schroeder (1995), Pokorski (2000) and Weinberg (1995) have excellent treatments of different aspects of anomalies. The string textbook of Green *et al.* (1987) provides a good introduction to anomalies in higher dimensions. One of the best introductions to the physics of instantons is provided in the lecture by Coleman (1985). The $U(1)$ problem in two-dimensional electrodynamics, and its role as a model for confinement, is discussed by Casher *et al.* (1974). The serious reader should study 't Hooft's instanton paper from 1976, in which he both uncovers much of the physical significance of the instanton solution, and performs a detailed evaluation of the determinant. Propagators in the instanton background are obtained in Brown *et al.* (1978).

Instantons in CP^N models are studied by Affleck (1980). The dependence of d_n on θ is calculated by Crewther *et al.* (1979) in a short and quite readable paper.

Exercises

(1) Derive Eq. (5.15).

(2) Calculate the decay rate of the π^0 to two photons. You will need the matrix element

$$\langle \pi(q) | \partial_\mu j_5^{\mu 3} | 0 \rangle = f_\pi q^\mu e^{iq\cdot x}, \tag{5.133}$$

where $f_\pi = 93\,\mathrm{MeV}$. You will need also to compute the anomaly in the third component of the axial isospin current.

(3) Fill in the details of the anomaly computation in two dimensions, being careful about signs and factors of 2.

(4) Fill in the details of the Fujikawa computation of the anomaly, in the CP^N model, being careful about factors of 2. Make sure you understand why one is calculating a determinant, and why the factors appear in the exponential. Verify that the action of Eq. (5.56) is equal to

$$\mathcal{L} = g_{\phi,\phi^*} \partial_\mu \phi \partial_\mu \phi^*, \tag{5.134}$$

where g is the metric of the sphere in complex coordinates, i.e. it is the line element $dx_1^2 + dx_2^2 + dx_3^2$ expressed as $g_{z,z} dz\, dz + g_{z,z^*} dz\, dz^* + g_{z^*z} dz^*\, dz + g_{dz^* dz^*} dz^* dz^*$. A model with an action of this form is called a "Non-linear Sigma Model;" the idea is that the fields live on some "target" space, with metric g. Verify Eqs. (5.56) and (5.59).

(5) Check that Eqs. (5.86) and (5.87) solve (5.83).

6

Grand unification

One of the troubling features of the Standard Model is the plethora of coupling constants; overall, there are 18, counting θ. It seems puzzling that a theory which purports to be a fundamental theory should have so many parameters. Another is the puzzle of charge quantization: why are the hypercharges all rational multiplets of one another (and, as a result, the electric charges are rational multiples of one another)? Finally, the gauge group itself is rather puzzling. Why is it semi-simple and not simple?

Georgi and Glashow put forward a proposal which answers some of these questions. They suggested that the underlying gauge symmetry of nature is a simple group, broken at some high-energy scale down to the gauge group of the Standard Model. The Standard Model gauge group has rank 4 (there are four commuting generators). $SU(N)$ groups have rank $N-1$. So the simplest group among the $SU(N)$s which might incorporate the Standard Model is $SU(5)$. Without any fancy group theory, it is easy to see how to embed $SU(3) \times SU(2) \times U(1)$ in $SU(5)$. Consider the gauge bosons. These are in the adjoint representation of the group. Written as matrices, under infinitesimal space-time independent gauge transformations,

$$\delta A_\mu = i\omega^a [T^a, A_\mu]. \tag{6.1}$$

The T_as are 5×5, traceless, Hermitian matrices; altogether, there are 24 of them. We can then break up the gauge generators in the following way. Writing the indices on T^a as $(T^a)_i^j$, the T^as act on the fundamental 5 representation as:

$$(T^a)_i^j 5_j. \tag{6.2}$$

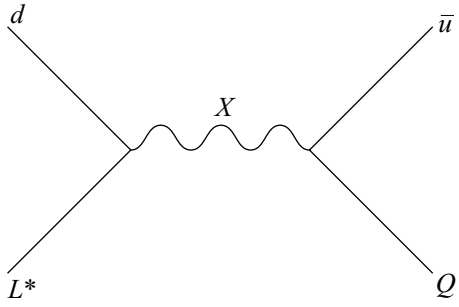

Fig. 6.1. Exchange of heavy vector particles in GUTs violates B and L. They can lead to processes such as $p \to \pi^0 e^+$.

So if we think of the 5 as:

$$5 = \begin{pmatrix} q_1 \\ q_2 \\ q_3 \\ L_1 \\ L_2 \end{pmatrix} \tag{6.3}$$

then the T^as can be broken up into a set of $SU(3)$ generators and a set of $SU(2)$ generators:

$$T^a = \begin{pmatrix} \frac{\lambda^a}{2} & 0 \\ 0 & 0 \end{pmatrix} \quad T^i = \begin{pmatrix} 0 & 0 \\ 0 & \frac{\sigma^i}{2} \end{pmatrix}. \tag{6.4}$$

Here the λ^as are Gell-Mann's $SU(3)$ matrices, and the σ^is are the Pauli matrices. There are three commuting matrices among these. The remaining diagonal matrix can be taken to be

$$\tilde{Y} = \frac{1}{\sqrt{60}} \begin{pmatrix} -2 & 0 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix}. \tag{6.5}$$

Finally, there are twelve off-diagonal matrices:

$$\left(X_a^i\right)_j^b = \delta_j^i \delta_a^b \tag{6.6}$$

where $a, b = 1, 2, 3$; $i, j = 1, 2$. These are not Hermitian; they are analogous to raising and lowering operators in $SU(2)$. One can readily form Hermitian linear combinations. The associated vector mesons must be very heavy; they mediate B-violating processes, as in Fig. 6.1. These can lead, for example, to $p \to \pi^0 e^+$.

We want to claim that $\tilde{Y}$ is proportional to the ordinary hypercharge, and determine the proportionality constant. To do this, we consider, not the 5, but the $\bar{5}$, and identify:

$$\bar{5} = \begin{pmatrix} \bar{d}_1 \\ \bar{d}_2 \\ \bar{d}_3 \\ L_1 \\ L_2 \end{pmatrix}. \tag{6.7}$$

Now the generators of $SU(5)$ acting on the $\bar{5}$ are $-T^{aT}$. So we can read off immediately that $Y = \sqrt{60}\tilde{Y}/3$. Since the gauge groups are unified in a single group, the gauge couplings are all the same, so we can compute the Weinberg angle. Calling g the $SU(5)$ coupling,

$$g\tilde{Y} = \frac{g'}{2}Y \tag{6.8}$$

where g' is the hypercharge coupling of the Standard Model. From this, $g^2 = 5/3{g'}^2$. The Weinberg angle,

$$\sin^2(\theta_W) = \frac{{g'}^2}{g^2 + {g'}^2} = \frac{3}{8}. \tag{6.9}$$

So we have two dramatic predictions, if we assume that the Standard Model is unified in this way.

(1) The $SU(3)$ and $SU(2)$ gauge couplings are equal.
(2) The Weinberg angle satisfies $\sin^2(\theta_W) = 3/8$.

Before assessing these predictions, let's first figure out where we would put the rest of the quarks and leptons. In a single generation of the Standard Model, there are 15 fields. $SU(5)$ has a 10 representation, the antisymmetric product of two 5s. It can be written as an antisymmetric matrix, 10_{ij}. If i and j are both $SU(3)$ indices, we obtain a $(\bar{3}, 1)_{-4/3}$ of $SU(3)$. If one is an $SU(3)$, one an $SU(2)$ index, we obtain a $(3, 2)_{1/3}$. If both are $SU(2)$ indices, we obtain a $(1, 1)_2$. Here the subscripts denote the ordinary hypercharge, related to $\tilde{Y}$ as above. These are just the quantum numbers of the Q, $\bar{u}$ and $\bar{e}$. As a matrix,

$$10 = \begin{pmatrix} 0 & \bar{u}^3 & -\bar{u}^2 & Q_1^1 & Q_1^2 \\ -\bar{u}^3 & 0 & \bar{u}^1 & Q_2^1 & Q_2^2 \\ \bar{u}^2 & -\bar{u}^1 & 0 & Q_3^1 & Q_3^2 \\ -Q_1^1 & -Q_2^1 & -Q_3^1 & 0 & \bar{e} \\ -Q_1^2 & -Q_2^2 & -Q_3^2 & -\bar{e} & 0 \end{pmatrix}. \tag{6.10}$$

So a single generation of quarks and leptons fits neatly into a $\bar{5}$ and 10 of $SU(5)$.

6.1 Cancellation of anomalies

An anomaly in a gauge symmetry would represent a breakdown of gauge invariance. The consistency of gauge symmetries rests, however, on gauge invariance. For example, to demonstrate that the theories are both unitary and Lorentz invariant, we have used different gauges. Cancellation of anomalies is crucial, and the absence of anomalies in the Standard Model is surely no accident.

It is not hard to check that in $SU(5)$, the anomaly of the $\bar{5}$ cancels that of the 10. In general, the anomalies in a gauge theory are proportional to d_{abc}, where

$$\{T_a, T_b\} = d_{abc} T_c. \tag{6.11}$$

One can organize the anticommutator above in terms of the various types of generators, for example $SU(3)$, $SU(2)$, $U(1)$, and the off-diagonal generators, which transform as (3, 2) of $SU(3) \times SU(2)$), and check each class. We leave the details for the exercises.

6.2 Renormalization of couplings

If we are going to describe the Standard Model, $SU(5)$ must break at some high-energy scale to $SU(3) \times SU(2) \times U(1)$. Above this scale, the full $SU(5)$ symmetry holds to a good approximation, and all couplings renormalize the same way. Below this scale, the couplings renormalize differently. We can write the equations for the renormalization of the three separate couplings:

$$\alpha_i^{-1}(\mu) = \alpha_{\rm gut}^{-1}(M_{\rm gut}) + \frac{b_0^i}{4\pi} \ln\left(\frac{\mu}{M_{\rm gut}}\right). \tag{6.12}$$

We can calculate the beta functions at one loop starting with the usual formula:

$$b_0 = \frac{11}{3} C_A - \frac{4}{3} c_{\rm f}^{(i)} N_{\rm f}^{(i)} - \frac{1}{3} c_\phi^{(i)} N_\phi^{(i)}, \tag{6.13}$$

where $N_{\rm f}^{(i)}$ is the number of fermions in the ith representation; $N_\phi^{(i)}$ is the number of scalars. For $SU(N)$, $C_{\rm A} = N$, and for fermions or scalars in the fundamental representation, $c_{\rm f} = c_\phi = 1/2$.

For $SU(3)$ and $SU(2)$ the beta-function coefficients, b_0^i, are readily computed. For the $U(1)$, we need to remember the relative normalization we computed above:

$$b_0^2 = 181/6 \quad b_0^3 = 7 \quad b_0^1 = \frac{61}{15}. \tag{6.14}$$

We can run these equations backwards. The $SU(2)$ and $U(1)$ couplings are the best measured, so it makes sense to start with these, and run them up to the unification scale. This determines $\alpha_{\rm gut}$ and $M_{\rm gut}$. We can then predict the value of the $SU(3)$ coupling at, say, M_Z. One finds that the unification scale, $M_{\rm gut}$, is about 10^{15} GeV, and α_3 is off by about seven standard deviations. In the exercises, you will have the opportunity to perform this calculation in detail. We will see later that low-energy supersymmetry greatly improves this.

6.3 Breaking to $SU(3) \times SU(2) \times U(1)$

In $SU(5)$, it is relatively easy to introduce a set of Higgs fields which break the gauge symmetry down to $SU(3) \times SU(2) \times U(1)$. Consider a Hermitian scalar field, Φ, in the adjoint representation. Writing Φ as a matrix, we have the transformation law:

$$\delta\Phi = \omega^a [T^a, \Phi]. \tag{6.15}$$

Suppose that the minimum of the Φ potential lies at a point where:

$$\Phi = v\tilde{Y}. \tag{6.16}$$

Then the $SU(3)$, $SU(2)$ and $U(1)$ generators all commute with $\langle\Phi\rangle$, but the Xs do not.

Consider the most general $SU(5)$-invariant potential:

$$V = -m^2 {\rm Tr}\Phi^2 + \lambda {\rm Tr}\Phi^4 + \lambda' ({\rm Tr}\Phi^2)^2. \tag{6.17}$$

One can find the minimum of this potential by first using an $SU(5)$ transformation to diagonalize Φ,

$$\Phi = {\rm diag}(a_1, a_2, a_3, a_4, a_5). \tag{6.18}$$

The potential is a function of the a_is, which one wants to minimize subject to the constraint of vanishing trace. This can be done by using a Lagrange multiplier.

To establish that one has a local minimum of the form of Eq. (6.16), one can proceed more simply. Write the potential as a function of v:

$$V = -\frac{1}{2}\mu^2 v^2 + \frac{a\lambda}{4} + \frac{b\lambda'}{4} v^4 \tag{6.19}$$

where $a = 7/120$. Then the extremum with respect to v is:

$$v = \frac{\mu}{\sqrt{a\lambda + b\lambda'}}. \tag{6.20}$$

To establish that this is a local minimum, we need to show that the eigenvalues of the scalar mass-squared matrix are all positive. We can investigate this

by considering small fluctuations about the stationary point. This point preserves $SU(3) \times SU(2) \times U(1)$. Writing $\Phi = \langle\Phi\rangle + \delta\Phi$, $\delta\Phi$ can be decomposed under $SU(3) \times SU(2) \times U(1)$ as

$$\delta\Phi = (1, 1) + (8, 1) + (1, 3) + (3, 2) + (\bar{3}, 2). \tag{6.21}$$

This point is certainly stationary; because of the symmetry, only the (1, 1) piece can appear linearly in the potential, and it is this piece whose minimum we have just found. To establish that the point is in fact a local minimum, one needs to show that the quadratic terms in the fluctuations are all positive. This is done in the exercises.

6.4 $SU(2) \times U(1)$ breaking

In addition to the adjoint, it is necessary to include a 5 of Higgs, H, in order to break $SU(2) \times U(1)$ to the $U(1)$ of electromagnetism and to give mass to the quarks and leptons. H has the form

$$H = \begin{pmatrix} H_{\rm c} \\ H_{\rm d} \end{pmatrix}, \tag{6.22}$$

where $H_{\rm c}$ is a color triplet of scalars and $H_{\rm d}$ is the ordinary Higgs doublet. For H one might be tempted to write a potential of the form

$$V(H) = -\mu^2|H|^2 + \frac{\lambda}{4}|H|^4. \tag{6.23}$$

However, this leads to a number of difficulties. Perhaps most important, when included in the larger theory with the adjoint field, Φ, this potential has too much symmetry; there is an extra $SU(5)$, which would lead to an assortment of unwanted Goldstone bosons. At the same time, the scale, μ, must be of order the scale of electroweak symmetry breaking (as long as λ is not too much larger than one). So the Higgs triplets will have masses of order the weak scale. But if the doublet couples to quarks and leptons, the triplet will have *baryon and lepton number violating* couplings to the quarks and leptons. So the triplet must be very massive.

Both problems can be solved if we couple Φ to H. The allowed couplings include:

$$V_{\Phi-H} = \Gamma H^*\Phi H + \lambda' H^* H \mathrm{Tr}\,\Phi^2 + \lambda'' H^*\Phi^2 H. \tag{6.24}$$

If we carefully adjust the constants Γ, λ', λ'' and μ^2, we can arrange that the doublets are light and the triplets are heavy. For example, if we choose $\lambda = \lambda' = 0$, and $\mu^2 = -3(\Gamma/\sqrt{60})v - \epsilon$ then the Higgs doublets have mass-squared $-\epsilon$ in the Lagrangian, while the triplets have mass of order $M_{\rm gut}$. This tuning of parameters, which must be performed in each order of perturbation theory, provides an explicit realization of the hierarchy problem.

Turning to fermion masses, we are led to an interesting realization: not only does grand unification make predictions for the gauge couplings, it can predict relations among fermion masses as well. $SU(5)$ permits the following couplings:

$$\mathcal{L}_y = y_1 \epsilon_{ijklm} H^i 10^{jk} 10^{lm} + y_2 H_i^* \bar{5}_j 10^{ij}. \tag{6.25}$$

Here the ys are matrices in the space of generations. When H acquires an expectation value, it gives mass to the quarks and leptons. The first coupling gives mass to the up-type quarks. The second coupling gives mass to both the down-type quarks and the leptons. If we consider only the heaviest generation, we then have the tree level prediction:

$$m_b = m_\tau. \tag{6.26}$$

This "prediction" is off by a factor of 3, but like the prediction of the coupling constant, it is corrected by renormalization by roughly the observed amount. For the lightest quarks and leptons, the prediction fails. However, unlike the unification of gauge couplings, such predictions can be modified if there are additional Higgs fields in other representations. In addition, for the lightest fermions, higher-dimension operators, suppressed by powers of the Planck mass, can make significant contributions to masses. In supersymmetric grand unified theories, the ratio of the GUT scale to the Planck scale is about 10^{-2}, whereas the lightest quarks and leptons have masses four orders of magnitude below the weak scale. We will postpone numerical study of these corrections, since the simplest $SU(5)$ theory does not correctly predict the values of the coupling constants. We will return to this subject when we discuss supersymmetric grand unified theories, which do successfully predict the observed values of the couplings.

6.5 Charge quantization and magnetic monopoles

While we must postpone success with the calculation of the unified couplings to our chapters on supersymmetry, we should pause and note two triumphs. First, we have a possible explanation for one of physics' greatest mysteries: why is electric charge quantized? Here it is automatic; electric charge, an $SU(5)$ generator, is quantized, just as color and isospin are quantized.

But Dirac long ago offered another explanation of electric charge quantization: magnetic monopoles. He realized that consistency of quantum mechanics demands that even if a single monopole exists in the universe, electric charges must all be integer multiples of a fundamental charge. So we might suspect that magnetic monopoles are hidden somewhere in this story. Indeed they are; these topics are discussed in Chapter 7.

6.6 Proton decay

We have discussed the dimension-six operators which can arise in the Standard Model and violate baryon number. Exchanges of the X bosons generate operators such as:

$$\frac{g^2}{M_X^2} Q\sigma_\mu \bar{u}^* Q\sigma^\mu \bar{e}^*. \tag{6.27}$$

This leads to the decay $p \to \pi^0 e^+$. In this model, one predicts a proton lifetime of order 10^{28} years if $M_{\rm gut} \approx 10^{15}$ GeV. The current limit on this decay mode is 5×10^{33} years. We will discuss the situation in supersymmetric models later.

The realization that baryon number violation is likely in any more fundamental theory opens up a vista on a fundamental question about nature: why is there more matter than antimatter in the universe. If, at some very early time, there were equal amounts of matter and antimatter, then if baryon number is violated, one has the possibility of producing an excess. Other conditions must be satisfied as well; we will describe this in the chapter on cosmology.

6.7 Other groups

While $SU(5)$ may in some respects be the simplest group for unification, once one has set off in this direction, there are many possibilities. Perhaps the next simplest is unification in the group $O(10)$. As $O(10)$ has rank 5, there is one extra commuting generator; presumably this symmetry must be broken at some scale. But more interesting is the fact that a single generation fits neatly in an irreducible representation: the 16. The $O(10)$ has an $SU(5)$ subgroup, under which the 16 decomposes as a $10 + \bar{5} + 1$. The singlet has precisely the right Standard Model quantum numbers – none – to play the role of the right-handed neutrino in the seesaw mechanism.

We won't thoroughly review the group theory of O groups, but we can describe some of the important features. We will focus specifically on $O(10)$, but much of the discussion here is easily generalized to other groups. The generators of $O(10)$ are 10×10 antisymmetric matrices. There are 45 of these. We are particularly interested in how these transform under the Standard Model group. The embedding of the Standard Model in $SU(5)$, as we have learned, is very simple, so a useful way to proceed to understand $O(10)$ is to find its $SU(5)$ subgroup.

One way to think of $O(10)$ is as the group of rotations of ten-dimensional vectors. Call the components of such a vector x^A, $A = 1, \ldots, 10$. $SU(5)$ transformations are "rotations" of complex five-dimensional vectors, z^i. So define

$$z^1 = x^1 + ix^2 \quad z^2 = x^3 + ix^4 \quad z^3 = x^5 + ix^6 \tag{6.28}$$

and so on. With this correspondence, it is easy to see that a subgroup of $O(10)$ transformations preserves the product $z \cdot z'^*$. This is the $SU(5)$ subgroup of $O(10)$.

From our construction, it follows that the 10 of $O(10)$ transforms as a $5 + \bar{5}$ of $SU(5)$. We can determine the decomposition of the adjoint by writing:

$$A^{AB} = A^{i\bar{i}} + A^{ij} + A^{\bar{i}\bar{j}}. \tag{6.29}$$

The labeling here is meant to indicate the types of complex indices the matrix A can carry. The first term is just the 24 of $SU(5)$, plus an additional singlet. This singlet is associated with a $U(1)$ subgroup of $O(10)$, which rotates all of the objects with i-type indices by one phase, and all of those with $\bar{i}$ type indices by the opposite phase. Note that A^{ij} is antisymmetric in its indices; in our study of $SU(5)$, we learned that this is the 10 representation. We can take it to carry charge 2 under the $U(1)$ subgroup. $A^{\bar{i}\bar{j}}$, then, is the $\overline{10}$ representation, with charge -2. This accounts for all 45 fields.

But where is the 16? We are familiar, from our experience with ordinary rotations in 3 and (Euclidean 4) dimensions, as well as from the Lorentz group, with the fact that O groups may have spinor representations. To construct these, we need to introduce the equivalent of Dirac gamma matrices, Γ, satisfying:

$$\{\Gamma^I, \Gamma^J\} = 2\delta^{IJ}. \tag{6.30}$$

It is not hard to construct explicit matrices which satisfy these anticommutation relations, but there is a simpler approach, which also makes the $SU(5)$ embedding clear. The anticommutation relations are similar to the relations for fermion creation and annihilation operators. So define:

$$a^1 = \frac{1}{2}(\Gamma^1 + i\Gamma^2) \quad a^2 = \frac{1}{2}(\Gamma^3 + i\Gamma^4) \tag{6.31}$$

and so on, and similarly for their complex conjugates. Note that the a^is form a 5 of $SU(5)$, with charge $+1$ under the $U(1)$. These operators satisfy the algebra:

$$\{a^i, a^{\bar{j}}\} = \delta^{i\bar{j}}. \tag{6.32}$$

These are the anticommutation relations for five pairs of fermion creation–annihilation operators. We know how to construct the corresponding "states," i.e. the representations of the algebra. We define a state, $|0\rangle$, annihilated by the a^is. Then there are five states created by $a^{\bar{i}}$ acting on this state:

$$\bar{5}_{-1} = a^{\bar{i}}|0\rangle. \tag{6.33}$$

The $\bar{5}$ indicates the $SU(5)$ representation; the subscript the $U(1)$ charge. We could now construct the states obtained with two creation operators, but let's construct

the states built with an odd number:

$$10_{-3} = a^{\bar{i}} a^{\bar{j}} a^{\bar{k}} |0\rangle \quad 1_{-5} = a^{\bar{1}} a^{\bar{2}} a^{\bar{3}} a^{\bar{4}} a^{\bar{5}} |0\rangle. \tag{6.34}$$

We have indicated that the first representation transforms like a 10 of $SU(5)$, while the second transforms like a singlet.

The states which involve even numbers of creation operators transform like a 5, a $\overline{10}$, and a singlet. Why do we distinguish these two sets? Remember, the goal of this construction is to obtain *irreducible* representations of the group $O(10)$. As for the Dirac theory, we can construct the symmetry generators from the Dirac matrices,

$$S^{IJ} = \frac{i}{4}[\Gamma^I, \Gamma^J]. \tag{6.35}$$

These, too, can be decomposed on the complex basis, like A^{IJ}. But, as for the usual Dirac matrices, there is another Γ matrix we can construct, the analog of Γ^5, Γ^{11}. This matrix anticommutes with all of the Γs, and thus the a^is. Thus the states with even numbers of creation operators are eigenstates with eigenvalue $+1$ under Γ^{11}, while those with odd numbers are eigenstates with eigenvalue -1. Since Γ^{11} commutes with the symmetry generators, these two representations are irreducible.

A similar construction works for other groups. When we come to discuss string theories in ten dimensions, we will be especially interested in the representations of $O(8)$. Here the same construction yields two representations, denoted by 8 and $8'$.

The embedding of the states of the Standard Model in $O(10)$ is clear, since we already know how to embed them in a $\bar{5} + 10$ of $SU(5)$. But what of the other state in the 16? This is a Standard Model singlet. We don't have a candidate, as of yet, in the particle data book for this. But there are two observations we can make. First, the symmetries of the Standard Model do not forbid a mass for this particle. What does forbid a mass is the extra $U(1)$. So if this symmetry is broken at very high energies, perhaps with the initial breaking of the gauge symmetry, this particle can gain a large mass. We will not explore the possible Higgs fields in $O(10)$, but, as in $SU(5)$, there are many possibilities, and the $U(1)$ can readily be broken. Second, this particle has the right quantum numbers to couple to the left-handed neutrino of the Standard Model. So this particle can naturally lead to a "seesaw" neutrino mass. This mass might be expected to be of order some typical Yukawa coupling, squared, divided by the unification scale. It is also possible that this extra $U(1)$ is broken at some lower scale, yielding a larger value for the neutrino mass.

Suggested reading

There are any number of good books and reviews on the subject of grand unification. The books by Ross (1984), Mohapatra (2003) and Ramond (1999) all treat the topics introduced in this chapter in great detail. The reader will find his or her interest increases after studying some aspects of supersymmetry.

Exercises

(1) Verify the cancellation of anomalies between the $\bar{5}$ and 10 representations of $SU(5)$.
(2) Establish the conditions that the solution of Eq. (6.16) is a local minimum of the potential.
(3) Perform the calculation of coupling unification in the $SU(5)$ model. Verify Eq. (6.14) for the $SU(3)$, $SU(2)$ and $U(1)$ beta functions. Start with the measured values of the $SU(2)$ and $U(1)$ couplings, being careful about the differing normalizations in the Standard Model and in $SU(5)$. Compute the value of the unification scale (the point where these two couplings are equal); then determine the value of α_3 at M_Z. Compare with the value given by the Particle Data Group. You need only study the equations to one-loop order. In practice, two-loop corrections, as well as threshold effects and higher-order corrections to the beta function, are often included.
(4) Add to the Higgs sector of the $SU(5)$ theory a set of scalars in the 45 representation. Show that in this case all of the quark masses are free parameters.

7

Magnetic monopoles and solitons

Anyone who has stared even briefly at Maxwell's equations has speculated about the existence of magnetic monopoles. There is no experimental evidence for magnetic monopoles, but the equations would be far more symmetric if they existed. It was Dirac who first considered carefully the implications of monopoles, and he came to a striking conclusion: the existence of monopoles would require that *electric charge* be quantized in terms of a fundamental unit. The problem of describing a monopole lies in writing $\vec{B} = \vec{\nabla} \times \vec{A}$. We could simply give up this identification, but Dirac recognized that $\vec{A}$ is essential in formulating quantum mechanics. To resolve the problem, we can follow Wu and Yang, and maintain $\vec{B} = \vec{\nabla} \times \vec{A}$, but not require that the vector potential be single valued. Suppose we have a monopole located at the origin. In the northern hemisphere, we can take

$$A_{\mathrm{N}} = \frac{g}{4\pi r} \frac{(1 - \cos(\theta))}{\sin(\theta)} \hat{e}_{\phi}, \tag{7.1}$$

while in the southern hemisphere we can take:

$$A_{\mathrm{S}} = -\frac{g}{4\pi r} \frac{(1 + \cos(\theta))}{\sin(\theta)} \hat{e}_{\phi}. \tag{7.2}$$

By looking up the formulae for the curl in spherical coordinates, you can check that, in both hemispheres:

$$\vec{B} = \frac{g}{4\pi r^2} \hat{r}, \tag{7.3}$$

so indeed this does describe a magnetic monopole.

Each of these expressions is singular along a half-line: A_{N} is singular along $\theta = \pi$; A_{S} is singular along $\theta = 0$. These string-like singularities are known as *Dirac strings*. They are suitable vector potentials to describe infinitely long, thin solenoids, starting at the origin and going to infinity on the negative or positive z axis. With a discontinuous $\vec{A}$, though, we need to ask whether quantum mechanics

is consistent. Consider the equator ($\theta = \pi/2$). Here

$$\vec{A}_{\rm N} - \vec{A}_{\rm S} = \frac{g}{2\pi r}\hat{e}_\phi = -\vec{\nabla}\chi \quad \chi = -\frac{g}{2\pi}\phi. \tag{7.4}$$

So the difference has the form of a gauge transformation. But to be a gauge transformation, it must act sensibly on particles of definite charge. In particular, it must be single valued. As the particle circumnavigates the sphere, its wave function acquires a phase

$$e^{ie\int d\vec{x}\cdot\vec{A}}. \tag{7.5}$$

Potentially, this phase is different if we use $\vec{A}_{\rm N}$ or $\vec{A}_{\rm S}$, in which case the string is a detectable, real object. But the phases are the same if

$$\exp\left(i\frac{eg}{2\pi}\oint d\vec{x}\cdot\vec{\nabla}\phi\right) = 1 \quad \text{or} \quad eg = 2\pi n. \tag{7.6}$$

This is known as the *Dirac quantization condition*. Dirac argued that since e can be the charge of any charged particle, if there is even one monopole somewhere in the universe, this result shows that charge must be quantized.

In pure electrodynamics, the status of magnetic monopoles is obscure; the $\vec{B}$ field is singular and the energy is infinite. But in non-Abelian gauge theories with scalar fields (Higgs fields), monopoles often arise as finite-energy, non-dissipative solutions of the classical equations. Such solutions cannot arise in linear theories, like electrodynamics; all configurations in such a theory spread with time. Non-dissipative solutions can only arise in non-linear theories, and even then, such solutions – known as solitons – can only arise in special circumstances.

The simplest theory which exhibits monopole solutions is $SU(2)$ (more precisely $O(3)$) Yang–Mills theory with a single Higgs particle in the adjoint representation. But before considering this case, which is somewhat complicated, it is helpful to consider solitons in lower-dimensional situations.

7.1 Solitons in $1+1$ dimensions

Consider a quantum field theory in $1+1$ dimension, with

$$\mathcal{L} = \frac{1}{2}(\partial_\mu\phi)^2 - V(\phi). \tag{7.7}$$

Here

$$V(\phi) = -\frac{1}{2}m^2\phi^2 + \lambda\phi^4. \tag{7.8}$$

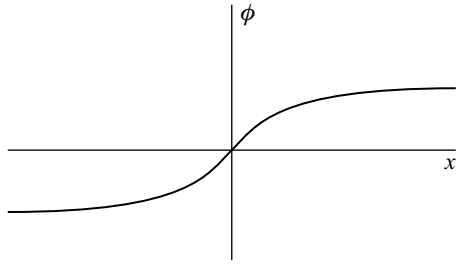

Fig. 7.1. Kink solution of the two-dimensional field theory.

This potential, which is symmetric under $\phi \to -\phi$, has two degenerate minima, $\pm\phi_0$. Normally, we would choose as our vacuum a state localized about one or the other minimum. These correspond to trivial solutions of the equations of motion. But we can consider a more interesting solution, for which

$$\phi(x \to \pm\infty) \to \pm\phi_0. \tag{7.9}$$

Such a solution interpolates between the two different vacua. We can construct this solution much as one solves analogous problems in classical mechanics, by quadrature. Finding the solution, known as the kink, is left for the exercises; the result is

$$\phi_{\text{kink}} = \phi_0 \tanh((x - x_0)m). \tag{7.10}$$

This solution is shown in Fig. 7.1. This object has finite energy. As we have indicated, there are a continuous infinity of solutions, corresponding to the fact that this *kink* can be located anywhere; this is a consequence of the underlying translational invariance. We can use this to understand in what sense the kink is a particle. Consider configurations which are not quite solutions of the equations of motion, in which x_0 is allowed to be a slowly varying function of t, $x_0(t)$. We can write the action for these configurations:

$$S_{\text{kink}} = \int dt \int dx \left[\frac{1}{2}(\partial_\mu \phi_{\text{kink}})^2 - V(\phi_{\text{kink}})\right]. \tag{7.11}$$

Only the $\dot{\phi}$ term contributes. The result is:

$$S_{\text{kink}} = \int dt \frac{M}{2}\dot{x}_0^2. \tag{7.12}$$

Here M is precisely the energy of the kink. So the kink truly acts as a particle. The x_0 is called a *collective coordinate*. We will see that such collective coordinates arise for each symmetry broken by the soliton. These are similar to the collective coordinates we encountered in the Euclidean problem of the instanton.

7.2 Solitons in $2+1$ dimensions: strings or vortices

As we go up in dimension, the possible solitons become more interesting. Consider a $U(1)$ gauge theory in $2+1$ dimensions, with a single charged scalar field, ϕ. This model is often called the Abelian Higgs model. The Lagrangian is:

$$\mathcal{L} = |D_\mu \phi|^2 - V(|\phi|). \tag{7.13}$$

We assume that the potential is such that

$$\langle \phi \rangle = v. \tag{7.14}$$

Now we have a possibility we haven't considered before. Working in plane polar coordinates, if we consider only the potential, we can imagine obtaining finite-energy configurations for which, at large r,

$$\phi \to e^{in\theta} v. \tag{7.15}$$

Because the potential tends to its minimum at infinity, such a configuration has finite potential energy. However, the kinetic energy diverges since $\partial_\mu \phi$ includes $(1/r)\partial_\theta \phi$. We can try to cancel this with a non-vanishing gauge field. At ∞, the scalar field is a gauge transformation of the constant configuration, so to achieve finite energy we want to gauge transform the gauge field as well:

$$A_\theta \to n \tag{7.16}$$

so $D_\mu \phi \to 1/r^2$, or faster. It is not hard to construct such solutions numerically. As for the kinks, these configurations have collective coordinates, corresponding to the two translational degrees of freedom and a rotational (or charge) degree of freedom.

We can take these configurations as configurations in a $3+1$-dimensional theory, which are constant with respect to z. Viewed in this way, these are vortices, or strings. One has collective coordinates corresponding to transverse motions of the string, $x_0(z,t)$, $y_0(z,t)$. These string configurations could be quite important in cosmology. Such a broken $U(1)$ theory could lead to the appearance of long strings, which could carry enormous amounts of energy. For a time, these were considered a possible origin of inhomogeneities leading to formation of galaxies, but the data now disfavors this possibility.

7.3 Magnetic monopoles

Dirac's argument shows that, in the presence of a monopole, electric charges are all multiples of a basic charge. This means that the $U(1)$ is effectively compact. So a natural place to look for monopoles is in gauge theories where the $U(1)$ is a

subgroup of a simple group. The $SU(5)$ grand unified theory was an example of this type, where electric charge is quantized.

We start, though, with the simplest example of this sort, an $SU(2)$ gauge theory with Higgs fields in the adjoint representation, ϕ^a. Such a theory was first considered by Georgi and Glashow as a model for weak interactions without neutral currents, and is known as the Georgi–Glashow model. An expectation value for ϕ, $\phi^3 = v$, or

$$\phi = \frac{v}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{7.17}$$

leaves an unbroken $U(1)$. The spectrum includes massive charged gauge bosons, $W^{\pm}$, and a massless gauge boson, which we will call the photon, γ. By analogy to the string/vortex solutions, we look for finite-energy at ∞

$$\phi \to gv. \tag{7.18}$$

In the $2+1$-dimensional case, we could think of the gauge transformation as a mapping from the space at infinity (topologically a circle) onto the gauge group (also a circle). In three dimensions, we want gauge transformations which map the two sphere, S_2, into the gauge group $SU(2)$. For example, we can take

$$g(\vec{x}) = i\frac{\hat{x}^i\sigma^i}{2}. \tag{7.19}$$

This suggests an "Ansatz" (guess) for a solution:

$$\phi^a = \hat{r}^a h(r) \quad A_i^a = -\epsilon_{ij}^a \frac{\hat{r}^j}{r} j(r). \tag{7.20}$$

This solution is very symmetric: it is invariant under a combined rotation in spin and isospin (rather similar to the sorts of symmetries of the instanton solution). Note that h and j satisfy coupled, non-linear equations, which, in general, must be solved numerically. We can see from the form of the action that the mass is of order $1/g^2$. In the next section we show that an analytic solution can be obtained in a particular limit.

We can write an elegant expression for the number of times $g(x)$ maps the sphere into the gauge group:

$$N = \frac{1}{4\pi}\int dS^i \epsilon_{ijk} \mathrm{Tr}(g\partial_j g \partial_k g). \tag{7.21}$$

In terms of the field, ϕ,

$$N = \frac{1}{8\pi v^3}\int \epsilon^{ijk}\epsilon^{abc}\partial_i\phi^a\partial_j\phi^b\partial_k\phi^c. \tag{7.22}$$

Finally, we need a definition of the magnetic charge. A natural choice is

$$\int d^3x \, \frac{1}{v} \, \partial_i \left(\phi^a B_i^a \right) = \frac{4\pi N}{e}.$$

Putting these statements together, we see that this solution, the *'t Hooft–Polyakov monopole*, has one Dirac unit of magnetic charge.

7.4 The BPS limit

Prasad and Sommerfield wrote down an exact monopole solution in the limit that $V = 0$. This limit seemed, originally, rather artificial, but we will see later that some supersymmetric field theories automatically have vanishing potential for a subset of fields. What simplifies the analysis in this limit is that the equations for the monopole, which are ordinarily second-order non-linear differential equations, become first-order equations. We will shortly understand this by thinking about supersymmetry. But first, we can derive this by looking directly at the potential for the gauge and scalar fields. We start by deriving a bound, the BPS bound, on the mass of a static field configuration. Again, we call the gauge coupling e, to avoid confusion with the magnetic charge g:

$$M_{\rm m} = \int d^3x \frac{1}{2} \left[\frac{1}{e^2} \vec{B}^a \cdot \vec{B}^a + (\vec{D}\Phi)^a \cdot (\vec{D}\Phi)^a \right]. \tag{7.23}$$

We can compare this with

$$\begin{aligned} A_\pm &= \int d^3x \left[\frac{1}{e} \vec{B}^a \pm (\vec{D}\Phi)^a \right]^2 \\ &= \frac{1}{2} \int d^3x \left[\frac{1}{e^2} \vec{B}^{a2} + (\vec{D}\Phi)^{a2} \right] \pm \int d^3x \frac{1}{e} \vec{B}^a (\vec{D}\Phi)^a. \end{aligned} \tag{7.24}$$

In the last term, we can integrate by parts. You can check that this works for both parts of the covariant derivative, i.e. this term becomes:

$$\frac{1}{e} \int d^3x (\vec{D} \cdot \vec{B})^a \Phi^a - \frac{1}{e} \int d^2a \Phi^a \hat{n} \cdot \vec{B}^a. \tag{7.25}$$

The first term vanishes by the Bianchi identity (the Yang–Mills generalization of the equation $\vec{\nabla} \cdot \vec{B} = 0$). The second term is v times what we have defined to be the monopole charge, g. So we have:

$$A_\pm = \int d^3x \left[\frac{1}{e^2} \vec{B}^a \pm (\vec{D}\Phi)^a \right]^2 = M_{\rm m} \pm \frac{vg}{e}. \tag{7.26}$$

The left-hand side of this equation is clearly greater than zero, so we have shown that

$$M_{\rm m} \geq \left|\frac{vg}{e}\right|. \tag{7.27}$$

This bound, known as the Bogomolny or BPS bound, is saturated when

$$\vec{B}^a = \pm\frac{1}{e}(\vec{D}\Phi)^a. \tag{7.28}$$

Note that, while we have spoken so far throughout this chapter about $SU(2)$, this result generalizes to any gauge group, with Higgs in the adjoint representation. But let's focus on $SU(2)$, and try to find a solution which satisfies the Bogomolny bound. As in the case of the 't Hooft–Polyakov monopole, it is convenient to write:

$$\Phi^a = \frac{\hat{r}^a}{er}H(evr) \quad A_i^a = -\epsilon_{ij}^a\frac{\hat{r}^j}{er}(1 - K(evr)). \tag{7.29}$$

Here we are using a dimensionless variable, $u = evr$, in terms of which the Hamiltonian scales simply. We are looking for solutions for which $H \to 0$ and $K \to 1$ as $r \to 0$. Otherwise, the solutions will be singular at the origin. At ∞, we want the configuration to look like a gauge transformation of the vacuum solution, so

$$K \to 0 \quad H \to evr \quad \text{as } r \to \infty. \tag{7.30}$$

We will leave the details to the exercises, but it is straightforward to show that these equations are solved by:

$$H(y) = y\coth(y) - 1 \quad K(y) = \frac{y}{\sinh(y)}. \tag{7.31}$$

The monopole mass is

$$M_{\rm m} = \frac{vg}{e} = \frac{2\pi v}{e^2}, \tag{7.32}$$

as predicted by the BPS formula.

7.5 Collective coordinates for the monopole solution

In lower-dimensional examples, we witnessed the emergence of collective coordinates, which described the translations and other collective motions of the solitons. In the case of the monopole, we have similar collective coordinates. Again, the solutions violate translational invariance. As a result, we can generate new solutions by replacing $\vec{x}$ by $\vec{x} - \vec{x}_0$. Now viewing x_0 as a slowly varying function of t, we obtain, as before, the action of a particle of mass $M_{\rm m}$.

There is another collective coordinate of the monopole solution, which has quite remarkable properties. In the monopole solution, charged fields are excited. So the monopole solution is not invariant under the $U(1)$ gauge transformations of electrodynamics. One might think that this is not important; after all, we have stressed that gauge transformations are not real symmetries, but instead represent a redundancy of the description of a system. But we need to be more precise. In interpreting Yang–Mills instantons, we worked in $A_0 = 0$ gauge. In this gauge, the important gauge transformations are time-independent gauge transformations, and these fall into two classes: large gauge transformations and small gauge transformations. The small gauge transformations are those which fall rapidly to zero at infinity, and physical states must be invariant under these. For large gauge transformations, this is not the case, and they can correspond to physically distinct configurations.

For the monopole configurations, the interesting gauge transformations are those which tend, at infinity, to a transformation in the unbroken $U(1)$ direction. For large r, this direction is determined by the direction of the Higgs fields. We must be careful about gauge fixing, so again we work in $A_0 = 0$ gauge. For our collective motion, we want to study gauge transformations in this direction, which vary slowly in time. It is important, however, that we remain in $A_0 = 0$ gauge, so the transformations we will study are not quite gauge transformations. Specifically, we consider

$$\delta A_i = D_i \chi(t) \Phi / v, \tag{7.33}$$

but we transform A_0 by

$$\delta A_0 = D_0(\chi \Phi)/v - \dot{\chi} \Phi / v \tag{7.34}$$

and, in order that the Gauss law constraint be satisfied, $\delta \Phi = 0$. The action for χ has the form:

$$S = \frac{C}{2e^2} \dot{\chi}^2. \tag{7.35}$$

Note that χ is bounded between 0 and 2π, i.e. it is an angular variable. Its conjugate variable is like an angular momentum; calling this Q, we have

$$Q = p_\chi = \frac{C}{e^2} \dot{\chi} \quad H = \frac{1}{2C} e^2 Q^2. \tag{7.36}$$

In the case of a BPS monopole, the constant C is $e^2 M_{\mathrm{m}}/(2v^2)$. So each monopole has a tower of charged excitations, with energies of order e^2 above the ground state. These excitations of the monopole about the ground state are known as *dyons*. The mass formula for these states has the form, in the case of a BPS monopole:

$$M = vg + vQ^2/g. \tag{7.37}$$

We will understand this better when we embed this structure in a supersymmetric field theory.

7.6 The Witten effect: the electric charge in the presence of θ

We have argued that in a $U(1)$ gauge theory, it is difficult to see the effects of θ. But in the presence of a monopole, a θ term has a dramatic effect, pointed out by Witten: the monopole acquires an electric charge, proportional to θ.

We can see this first in a heuristic way. We work in a gauge with non-zero A_0, and take all fields static. Then

$$\vec{E} = -\vec{\nabla} A_0 \quad \vec{B} = \frac{g}{4\pi}\frac{\vec{r}}{r^2} + \vec{\nabla} \times \vec{A}. \tag{7.38}$$

For such a configuration, the θ term,

$$\mathcal{L}_\theta = \frac{\theta e^2}{8\pi^2}\vec{E} \cdot \vec{B} \tag{7.39}$$

takes the form

$$\mathcal{L}_\theta = -\frac{\theta e^2 g}{32\pi^2}\int d^3 r A_0 \vec{\nabla} \cdot \frac{\hat{r}}{r^2} = -\frac{\theta e^2 g}{8\pi^2}\int d^3 r A_0 \delta(\vec{r}). \tag{7.40}$$

We started with a magnetic monopole at the origin, but we now also have an electric charge at the origin, $(\theta e^2 g)/(8\pi^2)$.

One might worry that in this analysis one is dealing with a singular field configuration, and in the non-Abelian case the configuration is non-singular. We can give a more precise argument. Let's go back to $A_0 = 0$ gauge. In this gauge, we can sensibly write down the canonical Hamiltonian. In the absence of θ, the conjugate momentum to $\vec{A}$ is $\vec{E}$. But in the presence of θ, there is an additional contribution:

$$\vec{\Pi} = -\frac{d\vec{A}}{dt} + \frac{\theta e^2}{8\pi^2}\vec{B}. \tag{7.41}$$

Now let's think about the invariance of states under small gauge transformations. For $\theta = 0$, we saw that the small gauge transformations, with gauge parameter ω, are generated by

$$Q_\omega = \int d^3 x \vec{\nabla}\omega \cdot \vec{E}. \tag{7.42}$$

An interesting set of large gauge transformations are those with $\omega^a = \lambda \Phi^a / v$. For these, if we integrate by parts, we obtain a term which vanishes by Gauss's law (Gauss's law is enforced by the invariance under small gauge transformations), and a surface term. This surface term gives the total $U(1)$ charge times λ. But we can think of this another way. For the low-lying excitations, multiplication by e^{iQ_ω}

is like shifting the dynamical variable χ by a constant, λ. In general, the wave functions for χ have the form $e^{iq\chi}$, where q is quantized. So the states pick up a phase $e^{iq\lambda}$. This is just the transformation of a state of charge q under the global gauge transformation with phase λ.

In the presence of θ, however, the operator which implements time-independent gauge transformations is modified. The $\vec{E}$ is replaced by the canonical momentum above. Now acting on states, the extra piece gives a factor $(g\theta)/(2\pi)$ in the exponent. Even states with $q = 0$ now pick up a phase, so there is an additional contribution to the charge:

$$Q = n_e e - \frac{\theta n_m}{2\pi}. \tag{7.43}$$

7.7 Electric–magnetic duality

Anyone who has ever seen Maxwell's equations has speculated on a possible duality between electricity and magnetism. If there were magnetic charges, these equations would take the form:

$$\vec{\nabla} \cdot \vec{E} = \rho_e \quad \vec{\nabla} \cdot \vec{B} = \rho_m, \tag{7.44}$$

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} + \vec{j}_m \quad \vec{\nabla} \times \vec{B} = \frac{\partial \vec{E}}{\partial t} + \vec{j}_e. \tag{7.45}$$

These equations retain their form if we, replace $\vec{E} \to -\vec{B}$, $\vec{B} \to \vec{E}$ and also let $\rho_e \to \rho_m$, $\rho_m \to -\rho_e$ (and similarly for the electric and magnetic currents).

Now that we have a framework for discussing magnetic charges, it is natural to ask whether some theories of electrodynamics really obey such a symmetry. In general, however, this is a difficult problem. We have just learned that electric and magnetic charges, when they both exist, obey a reciprocal relation, $g \propto 1/e$. From the point of view of quantum field theory, this means that exchanging electric and magnetic charges also means replacing the fundamental coupling by its inverse. In other words, if there is such a duality symmetry, *it relates a strongly coupled theory to a weakly coupled theory*. We don't know a great deal about strongly coupled gauge theories, so investigating the possibility of such a duality is a difficult problem. That such a symmetry might exist in theories of the type we have been discussing is not entirely crazy. For example, the monopole masses behave, at weak coupling, like $1/g^2$. So as the coupling becomes strong, these particles become light, even as the charged states become heavy. They have complicated quantum numbers (some of the monopole states are fermionic, for example).

Remarkably, there is a circumstance where such dualities can be studied: theories with more than one supersymmetry (in four dimensions). $N = 4$ supersymmetric

Yang–Mills theory turns out to exhibit an electric–magnetic duality. These theories will be discussed in Chapter 15. Crucial to verifying this duality will be a deeper understanding of the BPS condition, which will allow us to establish exact formulas for masses of certain particles, valid for all values of the coupling. These formulas will exhibit precisely the expected duality between electricity and magnetism.

Suggested reading

There are many excellent reviews and texts on monopoles. These include Coleman (1981) and Harvey (1996), and this chapter borrows ideas from both. You can find an introduction to the subject in Chapter 6 of Jackson's electrodynamics text (1999).

Exercises

(1) Verify that Eqs. (7.1) and (7.2) are those of infinitely long, thin solenoids ending at the origin.

(2) Find the kink solution of the $1 + 1$-dimensional model. Show that the collective coordinate action is

$$S = \int dt \frac{1}{2} M_{\text{kink}} \dot{x}_0^2.$$

(3) Verify that Eq. (7.31) solves the BPS equations.

8

Technicolor: a first attempt to explain hierarchies

In Chapter 5, we learned a great deal about the dynamics of quantum chromodynamics. In Section 4.5, we argued that the hierarchy problem is one of the puzzles of the Standard Model. The grand unified models of the previous chapter provided a quite stark realization of the hierarchy problem. In an $SU(5)$ grand unified model, we saw that it is necessary to carefully adjust the couplings in the Higgs potential in order that one obtain light doublets and heavy color triplet Higgs. This is already true at tree level; loop effects will correct these relations, requiring further delicate adjustments.

The first proposal to resolve this problem goes by the name "technicolor" and is the subject of this chapter. The technicolor hypothesis exploits our understanding of QCD dynamics. It elegantly explains the breaking of the electroweak symmetry. It has more difficulty accounting for the masses of the quarks and leptons, and simple versions seem incompatible with precision studies of the W and Z particles. In this chapter, we will introduce the basic features of the technicolor hypothesis. We will not attempt to review the many models that have been developed to try to address the difficulties of flavor and precision electroweak experiments. It is probably safe to say that, as of this writing, none is totally successful, nor are they terribly plausible. But it should be kept in mind that this may reflect the limitations of theorists; experiment may yet reveal that nature has chosen this path. In the second part of this book, we will argue that in string theory, ignoring phenomenological details, a technicolored solution to the hierarchy problem seems as likely as its main competitor, supersymmetry. Perhaps some reader of this book will realize what theorists have been missing.

Apart from the fact that technicolor might have something to do with Nature, this brief chapter will also provide an opportunity to develop a deeper understanding of the non-perturbative aspects of gauge theories.

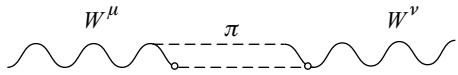

Fig. 8.1. Diagrammatic representation of technicolor.

8.1 QCD in a world without Higgs fields

Consider a world with only a single generation of quarks and no Higgs fields. In such a world, the quarks would be exactly massless. The $SU(2)_{\mathrm{L}} \times SU(2)_{\mathrm{R}}$ symmetry of QCD would be, in part, a gauge symmetry. $SU(2)_{\mathrm{L}}$ would correspond to the $SU(2)$ symmetry of the weak interactions. Hypercharge, Y, would include a generator of $SU(2)_{\mathrm{R}}$ and baryon number:

$$Y = 2T_{3\mathrm{R}} + B. \tag{8.1}$$

The quark condensate,

$$\langle q_f \bar{q}_{f'} \rangle = \Lambda^3 \delta_{ff'} \tag{8.2}$$

would break some of the gauge symmetry. Electric charge, however, would be conserved, so $SU(2) \times U(1) \to U(1)$.

In Appendix C, we saw that the quark condensate conserves a vector $SU(2)$, ordinary isospin. This $SU(2)$ is generated by the linear sum

$$T_{\mathrm{i}} = T_{\mathrm{iL}} + T_{\mathrm{iR}}. \tag{8.3}$$

So the $SU(2)$ gauge bosons transform as a triplet of the conserved isospin. This guarantees that the successful tree-level relation,

$$M_W = M_Z \cos(\theta), \tag{8.4}$$

is satisfied.

To understand the masses of the gauge bosons, remember that for a broken symmetry with current j^μ, the coupling of the Goldstone boson to the current is

$$\langle 0 | j^\mu | \pi(p) \rangle = i f_\pi p^\mu. \tag{8.5}$$

This means that there is a non-zero amplitude for a gauge boson to turn into a Goldstone, and vice versa. The diagram of Fig. 8.1 is proportional to:

$$g^2 f_\pi^2 p^\mu \frac{i}{p^2} p^\nu. \tag{8.6}$$

As the momentum tends to zero, this tends to a constant – the mass of the gauge boson. For the charged gauge bosons, the mass is just:

$$m_{W^\pm}^2 = g^2 f_\pi^2 \tag{8.7}$$

while for the neutral gauge bosons we have a mass matrix:

$$f_{\pi^2}\begin{pmatrix} g^2 & gg' \\ gg' & g'^2 \end{pmatrix} \tag{8.8}$$

giving one massless gauge boson and one with mass-squared $(g^2 + g'^2)f_\pi^2$.

All of this can be nicely described in terms of the non-linear sigma model which is used to describe pion physics. Recall that the pions could be described in terms of a matrix,

$$\Sigma = |\langle\overline{\psi}\psi\rangle| e^{i\frac{\vec{\pi}\cdot\vec{\tau}}{2}}, \tag{8.9}$$

which transforms under $SU(2)_{\mathrm{L}} \times SU(2)_{\mathrm{R}}$ as:

$$\Sigma \to U_{\mathrm{L}} \Sigma U_{\mathrm{R}}^{\dagger}. \tag{8.10}$$

Changes in the magnitude of the condensate are associated with excitations in QCD much more massive than the pion fields (the σ field of our linear sigma model of Section 2.2). So it is natural to treat this as a constant. Σ is then a field constrained to move on a manifold. As in our examples in two dimensions, a model based on such a field is called a non-linear sigma model. The Lagrangian is:

$$\mathcal{L} = f_\pi^2 \,\mathrm{Tr}\partial_\mu \Sigma^\dagger \partial^\mu \Sigma. \tag{8.11}$$

In the context of the physics of light pseudo-Goldstone particles, the virtue of such a model is it incorporates the effects of broken symmetry in a very simple way. For example, all of the results of current algebra can be derived by studying the physics of such a theory and its associated Lagrangian.

In the case of the σ-model, we have an identical structure, except that we have gauged some of the symmetry, so we need to replace the derivatives by covariant derivatives:

$$\partial_\mu \Sigma \to D_\mu \Sigma = \partial_\mu \Sigma - i\frac{A_\mu^a \sigma_a}{2}\Sigma - i\Sigma\frac{\sigma_3}{2}B_\mu. \tag{8.12}$$

Again, we can choose unitary gauge; we just set $\Sigma = 1$. So the Lagrangian in this gauge is simply:

$$\mathcal{L} = \mathrm{Tr}\left(\frac{A_\mu^a \sigma_a}{2}\Sigma + \Sigma\frac{\sigma_3}{2}B_\mu\right)^2. \tag{8.13}$$

This yields exactly the mass matrix we wrote before.

8.2 Fermion masses: extended technicolor

In technicolor models, the Higgs field is replaced by new strong interactions which break $SU(2) \times U(1)$ at a scale $F_\pi = 1$ TeV. But the Higgs field of the Standard

Model gives masses not only to the gauge bosons, but to the quarks and leptons as well. In the absence of the Higgs scalar, there are chiral symmetries which prohibit masses for any of the quarks and leptons. While our simple model can explain the masses of the Ws and Zs, it has no mechanism to generate the masses for the ordinary quarks and leptons.

If we are not to introduce fundamental scalars, the only way to break these symmetries is to introduce further gauge interactions. Consider first a single generation of quarks and leptons. Enlarge the gauge group to $SU(3) \times SU(2) \times U(1) \times SU(N+1)$. The technicolor group will be an $SU(N)$ subgroup of the last factor. Take each quark and lepton to be part of an $N+1$ or $\overline{N+1}$ of this larger group. To avoid anomalies, we will also include a right-handed neutrino. In other words, our multiplet structure is:

$$\begin{pmatrix} Q \\ q \end{pmatrix} \quad \begin{pmatrix} \bar{U} \\ \bar{u} \end{pmatrix} \quad \begin{pmatrix} \bar{D} \\ \bar{d} \end{pmatrix} \quad \begin{pmatrix} L \\ \ell \end{pmatrix} \quad \begin{pmatrix} \bar{E} \\ \bar{e} \end{pmatrix} \quad \begin{pmatrix} \bar{N} \\ \bar{\nu} \end{pmatrix}. \tag{8.14}$$

Here $q, \bar{u}, \bar{d}, \ell$, etc., are the usual quarks and leptons; the fields denoted with capital letters are the techniquarks. Now suppose that the $SU(N+1)$ is broken to $SU(N)$ by some other gauge interactions, in a manner similar to technicolor, at a scale $\Lambda_{\rm etc} \gg \Lambda_{\rm tc}$. Then there are a set of broken gauge generators with mass of order $\Lambda_{\rm etc}$. Exchanges of these bosons give rise to operators such as:

$$\mathcal{L}_{\rm 4f} = \frac{1}{\Lambda_{\rm etc}^2} Q \sigma_\mu q^* \bar{U} \sigma^\mu \bar{u}^* + \text{h.c.} \tag{8.15}$$

Using the identity for the Pauli matrices,

$$\sum_\mu (\sigma_\mu)_{\alpha\dot{\alpha}} (\sigma^\mu)^{\dot{\beta}\beta} = \delta_\alpha^\beta \delta_{\dot{\alpha}}^{\dot{\beta}}, \tag{8.16}$$

permits us to rewrite the four-fermi interaction:

$$\mathcal{L}_{\rm 4f} = \frac{1}{\Lambda_{\rm etc}} Q \bar{U} q^* \bar{u}^* + \text{h.c.} \tag{8.17}$$

We can replace $Q\bar{U}$ by its expectation value, of order $\Lambda_{\rm tc}^3$. This gives rise to a mass for the u quark. The other quarks and leptons gain mass in a similar fashion.

This model is clearly unrealistic on many counts: it has only one generation; there is a massive neutrino; there are relations among the masses which are unrealistic; there are approximate global symmetries which lead to unwanted pseudo-Goldstone bosons. Still, it illustrates the basic idea of extended technicolor models: additional gauge interactions break the unwanted chiral symmetries which protect the quark and lepton masses.

One can try to build realistic models by considering more complicated groups and representations for the extended technicolor (etc) interactions. Rather than attempt this here, we consider some of the issues in a general way. We imagine we have a model with three generations. The extended technicolor interactions generate a set of four-Fermi interactions which break the chiral symmetries which act on the separate quarks and leptons. In a model of three generations, there are a number of challenges which must be addressed.

(1) Perhaps the most serious is the problem of flavor-changing neutral currents. In addition to four-Fermi operators which generate mass, there will also be four-Fermi operators involving just the ordinary quarks and leptons. These operators will not, in general, respect flavor symmetries. They are likely to include terms like

$$\mathcal{L}_{\Delta S=2} = \frac{1}{\Lambda_{\text{etc}}} \bar{s}\bar{d}s^*d^*, \tag{8.18}$$

which violate strangeness by two units. Unless Λ_{etc} is extremely large (of the order of hundreds of TeV), this will lead to unacceptably large rates for $K^0 \leftrightarrow \bar{K}^0$.

(2) Generating the top quark mass is potentially problematic. The top quark mass is larger than the W and Z masses. Yet if the ETC scale is large, it is hard to see how to achieve this.

(3) The problem of pseudo-Goldstone bosons is generic to technicolor models, in just the fashion we saw for the simple model.

The challenge of technicolor model building is to construct models which solve these problems. We will not attempt to review the various approaches which have been put forward here. Models which solve these problems are typically extremely complicated. Instead, we will close by briefly discussing another serious difficulty: precision measurements of electroweak processes.

8.3 Precision electroweak measurements

In Section 4.5, we stressed that the parameters of the electroweak theory have been measured with high precision, and compared with detailed theoretical calculations, including radiative corrections. One naturally might wonder whether a strongly interacting Higgs sector could reproduce these results. The answer is that it is difficult. There are two categories of corrections which one needs to consider. The first are, in essence, corrections to the relation

$$M_W = M_Z \cos(\theta_W). \tag{8.19}$$

In a general technicolor model, these will be large. But we have seen why this relation holds in the minimal Standard Model: there is an approximate, global

$SU(2)$ symmetry. This is in fact the case of the simplest technicolor model we encountered above. So this problem likely has solutions.

There are, however, other corrections as well, resulting from the fact that in these strongly coupled theories, the gauge boson propagators are quite different than those of weakly coupled field theories. These have been estimated in many models, and are found to be far too large to be consistent with the data. More details about this problem, and speculations on possible solutions, can be found in the Suggested reading.

Suggested reading

An up-to-date set of lectures on technicolor, including the problems of flavor and electroweak precision measurements, are those of Chivukula. An introduction to the analysis of precision electroweak physics is provided by Peskin (1990); for the application to technicolor theories, see Peskin and Takeuchi (1990).

Exercises

(1) Determine the relations between the quark and lepton masses in the extended technicolor model above.

(2) What are the symmetries of the extended technicolor model in the limit that we turn off the ordinary $SU(3) \times SU(2) \times U(1)$ gauge interactions? How many of these symmetries are broken by the condensate? Each of these broken symmetries gives rise to an appropriate Nambu–Goldstone boson. Some of these approximate symmetries are broken explicitly by the ordinary gauge interactions. The corresponding Goldstone bosons will then gain mass, typically of order $\alpha_i \Lambda_{\rm etc}$. But some will not gain mass of this order. Which symmetry (or symmetries) will be respected by the ordinary gauge interactions?

Part 2

Supersymmetry

9

Supersymmetry

In a standard advanced field theory course, one learns about a number of symmetries: Poincaré invariance, global continuous symmetries, discrete symmetries, gauge symmetries, approximate and exact symmetries. These latter symmetries all have the property that they commute with Lorentz transformations, and in particular they commute with rotations. So the multiplets of the symmetries always contain particles of the same spin; in particular, they always consist of either bosons or fermions.

For a long time, it was believed that these were the only allowed types of symmetry; this statement was even embodied in a theorem, known as the Coleman–Mandula theorem. However, physicists studying theories based on strings stumbled on a symmetry which related fields of different spin. Others quickly worked out simple field theories with this new symmetry: supersymmetry.

Supersymmetric field theories can be formulated in dimensions up to eleven. These higher-dimensional theories will be important when we consider string theory. In this chapter, we consider theories in four dimensions. The supersymmetry charges, because they change spin, must themselves carry spin – they are spin-$1/2$ operators. They transform as doublets under the Lorentz group, just like the two-component spinors χ and χ^*. (The theory of two-component spinors is reviewed in Appendix A, where our notation, which is essentially that of the text by Wess and Bagger (1992), is explained). There can be 1, 2, 4 or 8 such spinors; correspondingly, the symmetry is said to be $N = 1, 2, 4$ or 8 supersymmetry. Like generators of an ordinary group, the supersymmetry generators obey an algebra; unlike an ordinary bosonic group, however, the algebra involves anticommutators as well as commutators (it is said to be "graded").

There are at least four reasons to think that supersymmetry might have something to do with TeV-scale physics. The first is the hierarchy problem: as we will see, supersymmetry can both explain how hierarchies arise, and why there are not large radiative corrections. The second is the unification of couplings. We have seen that

while the gauge group of the Standard Model can rather naturally be unified in a larger group, the couplings do not unify properly. In the minimal supersymmetric extension of the Standard Model (the Minimal Supersymmetric Standard Model, or MSSM) the couplings unify nicely, if the scale of supersymmetry breaking is about 1 TeV. Third, the assumption of TeV-scale supersymmetry almost automatically yields a suitable candidate for the dark matter, with a density in the required range. Finally, low-energy supersymmetry is strongly suggested by string theory, though at present one cannot assert that this is a prediction.

9.1 The supersymmetry algebra and its representations

Because the supersymmetry generators are spinors, they do not commute with the Lorentz generators. Perhaps, then, it is not surprising that the supersymmetry algebra involves the translation generators ($\bar{Q}_{\dot{\alpha}} = \bar{Q}^*_{\dot{\alpha}}$)

$$\left\{ Q^A_\alpha, \bar{Q}^B_{\dot{\beta}} \right\} = 2\sigma^\mu_{\alpha\dot{\beta}} \delta^{AB} P_{\mu s} \tag{9.1}$$

$$\left\{ Q^A_\alpha, \bar{Q}^B_\beta \right\} = \epsilon_{\alpha\beta} X^{AB}. \tag{9.2}$$

The X^{AB}s are Lorentz scalars, antisymmetric in A, B, known as central charges.

If nature is supersymmetric, it is likely that the low-energy symmetry is $N = 1$, corresponding to only one possible value for the index A above. Only $N = 1$ supersymmetry has chiral representations. Of course, one might imagine that the chiral matter would arise at the point where supersymmetry was broken. But, as we will see, it is very difficult to break $N > 1$ supersymmetry spontaneously; this is not the case for $N = 1$. The smallest irreducible representations of $N = 1$ supersymmetry which can describe massless fields are as follows.

- Chiral superfields: (ϕ, ψ_α), a complex scalar and a chiral fermion.
- Vector superfields: (λ, A_μ), a chiral fermion and a vector meson, both, in general, in the adjoint representation of the gauge group.
- The gravity supermultiplet: ($\psi_{\mu,\alpha}$, $g_{\mu\nu}$), a spin-3/2 particle, the gravitino, and the graviton.

One can work in terms of these fields, writing supersymmetry transformation laws and constructing invariants. This turns out to be rather complicated. One must use the equations of motion to realize the full algebra. Great simplification is achieved by enlarging space-time to include commuting and anticommuting variables. The resulting space is called *superspace*.

9.2 Superspace

We may conveniently describe $N = 1$ supersymmetric field theories by using superspace. Superspace allows a simple description of the action of the symmetry on

fields and provides an efficient algorithm for construction of invariant Lagrangians. In addition, calculations of Feynman graphs and other quantities are often greatly simplified using superspace, at least in the limit that supersymmetry is unbroken or nearly so.

In superspace, in addition to the ordinary coordinates, x^μ, one has a set of anticommuting, *Grassmann*, coordinates, θ_α, $\theta^*_{\dot\alpha} = \bar\theta_{\dot\alpha}$. The Grassmann coordinates obey:

$$\{\theta_\alpha, \theta_\beta\} = \{\bar\theta_{\dot\alpha}, \bar\theta_{\dot\beta}\} = \{\theta_\alpha, \bar\theta_{\dot\beta}\} = 0. \tag{9.3}$$

Grassmann coordinates provide a representation of the classical configuration space for fermions; they are familiar from the problem of formulating the fermion functional integral. Note that the square of any θ vanishes. Derivatives also anticommute:

$$\left\{\frac{\partial}{\partial\theta_\alpha}, \frac{\partial}{\partial\bar\theta_{\dot\beta}}\right\} = 0, \text{ etc.} \tag{9.4}$$

Crucial in the discussion of Grassmann variables is the problem of integration. In discussing Poincaré invariance of ordinary field theory Lagrangians, the property of ordinary integrals that

$$\int_{-\infty}^{\infty} dx f(x+a) = \int_{-\infty}^{\infty} dx f(x) \tag{9.5}$$

is important. We require that the analogous property hold for Grassmann integration (here for one variable):

$$\int d\theta f(\theta+\epsilon) = \int d\theta f(\theta). \tag{9.6}$$

This is satisfied by the integration rule:

$$\int d\theta(1, \theta) = (0, 1). \tag{9.7}$$

For the case of θ_α, $\bar\theta_{\dot\alpha}$, one can write a simple integral table:

$$\int d^2\theta\theta^2 = 1; \int d^2\bar\theta\bar\theta^2 = 1, \tag{9.8}$$

all others vanishing.

One can formulate a superspace description for both local and global supersymmetry. The local case is rather complicated, and we won't deal with it here, referring the interested reader to the suggested reading, and confining our attention to the global case.

The goal of the superspace formulation is to provide a classical description of the action of the symmetry on fields, just as one describes the action of the

Poincaré generators. Consider functions of the superspace variables, $f(x^\mu, \theta, \bar{\theta})$. The supersymmetry generators act on these as differential operators: x^μ, θ, θ^*:

$$Q_\alpha = \frac{\partial}{\partial\theta_\alpha} - i\sigma^\mu_{\alpha\dot{\alpha}}\theta^{*\dot{\alpha}}\partial_\mu; \quad \bar{Q}_{\dot{\alpha}} = -\frac{\partial}{\partial\bar{\theta}_{\dot{\alpha}}} + i\theta^{*\alpha}\sigma^\mu_{\alpha\dot{\alpha}}\partial_\mu. \tag{9.9}$$

Note that the θs have mass dimension $-1/2$. It is easy to check that the Q_αs obey the algebra. For example,

$$\{Q_\alpha, Q_\beta\} = \left\{\left(\frac{\partial}{\partial\theta_\alpha} - i\sigma^\mu_{\alpha\dot{\alpha}}\theta^{*\dot{\alpha}}\partial_\mu\right), \left(\frac{\partial}{\partial\theta_\beta} - i\sigma^\nu_{\beta\dot{\beta}}\theta^{*\dot{\beta}}\partial_\nu\right)\right\} = 0, \tag{9.10}$$

since the θs and their derivatives anticommute. With just slightly more effort, one can construct the $\{Q_\alpha, Q^*_{\dot{\alpha}}\}$ anticommutator.

One can think of the Qs as generating infinitesimal transformations in superspace with Grassmann parameter ϵ. One can construct finite transformations as well by exponentiating the Qs; because there are only a finite number of non-vanishing polynomials in the θs, these exponentials contain only a finite number of terms. The result can be expressed elegantly:

$$e^{\epsilon Q + \epsilon^* \bar{Q}}\Phi(x^\mu, \theta, \bar{\theta}) = \Phi(x^\mu - i\epsilon\sigma^\mu\theta^* + i\theta\sigma^\mu\epsilon^*, \theta + \epsilon, \bar{\theta} + \epsilon^*). \tag{9.11}$$

If one expands Φ in powers of θ, there are only a finite number of terms. These can be decomposed into two irreducible representations of the algebra, corresponding to the chiral and vector superfields described above. To understand these, we need to introduce one more set of objects, the covariant derivatives, D_α and $\bar{D}_{\dot{\alpha}}$. These are objects which anticommute with the supersymmetry generators, and thus are useful for writing down invariant expressions. They are given by

$$D_\alpha = \partial_\alpha + i\sigma^\mu_{\alpha\dot{\alpha}}\theta^{*\dot{\alpha}}\partial_\mu; \quad \bar{D}_{\dot{\alpha}} = -\partial_{\dot{\alpha}} - i\theta^\alpha\sigma^\mu_{\alpha\dot{\alpha}}\partial_\mu. \tag{9.12}$$

They satisfy the anticommutation relations:

$$\{D_\alpha, \bar{D}_{\dot{\alpha}}\} = -2i\sigma^\mu_{\alpha\dot{\alpha}}\partial_\mu \quad \{D_\alpha, D_\alpha\} = \{\bar{D}_{\dot{\alpha}}, \bar{D}_{\dot{\beta}}\} = 0. \tag{9.13}$$

We can use the Ds to construct irreducible representations of the supersymmetry algebra. Because the Ds anticommute with the Qs, the condition

$$\bar{D}_{\dot{\alpha}}\Phi = 0 \tag{9.14}$$

is invariant under supersymmetry transformations. Fields that satisfy this condition are called chiral superfields. To construct chiral superfields, we would like to find combinations of x^μ, θ and $\bar{\theta}$ which are annihilated by $\bar{D}_\alpha$. Writing

$$y^\mu = x^\mu + i\theta\sigma^\mu\bar{\theta}, \tag{9.15}$$

then

$$\Phi = \Phi(y) = \phi(y) + \sqrt{2}\theta\psi(y) + \theta^2 F(y) \tag{9.16}$$

is a chiral (scalar) superfield. Expanding in θ, the expansion terminates:

$$\Phi = \phi(x) + i\theta\sigma^\mu\bar{\theta}\partial_\mu\phi + \frac{1}{4}\theta^2\bar{\theta}^2\partial^2\phi \tag{9.17}$$
$$+ \sqrt{2}\theta\psi - \frac{i}{\sqrt{2}}\theta\theta\partial_\mu\psi\sigma^\mu\bar{\theta} + \theta^2 F.$$

We can work out the transformation laws. Starting with

$$\delta\Phi = \epsilon^\alpha Q_\alpha \Phi + \epsilon_\alpha \bar{Q}^{\dot{\alpha}} \tag{9.18}$$

the components transform as

$$\delta\phi = \sqrt{2}\epsilon\psi \quad \delta\psi = \sqrt{2}\epsilon F + \sqrt{2}i\sigma^\mu\epsilon^*\partial_\mu\phi \quad \delta F = i\sqrt{2}\epsilon^*\bar{\sigma}^\mu\partial_\mu\psi. \tag{9.19}$$

Vector superfields form another irreducible representation of the algebra; they satisfy the condition

$$V = V^\dagger. \tag{9.20}$$

Again, it is easy to check that this condition is preserved by supersymmetry transformations. V can be expanded in a power series in θs:

$$V = i\chi - i\chi^\dagger - \theta\sigma^\mu\theta^* A_\mu + i\theta^2\bar{\theta}\bar{\lambda} - i\bar{\theta}^2\theta\lambda + \frac{1}{2}\theta^2\bar{\theta}^2 D. \tag{9.21}$$

Here χ is not quite a chiral field; it is a superfield which is a function of θ only, i.e. it has terms with zero, one or two θs; χ^* is its conjugate.

If V is to describe a massless field, the presence of A_μ indicates that there should be some underlying gauge symmetry, which generalizes the conventional transformation of bosonic theories. In the case of a $U(1)$ theory, gauge transformations act by

$$V \to V + i\Lambda - i\Lambda^\dagger \tag{9.22}$$

where Λ is a chiral field. The $\theta\theta^*$ term in Λ is precisely a conventional gauge transformation of A_μ. In the case of a $U(1)$ theory, one can define a gauge-invariant field strength,

$$W_\alpha = -\frac{1}{4}\bar{D}^2 D_\alpha V. \tag{9.23}$$

By a gauge transformation, we can set $\chi = 0$. The resulting gauge is known as the Wess–Zumino gauge. This gauge is analogous to Coulomb gauge in

electrodynamics:

$$W_\alpha = -i\lambda_\alpha + \theta_\alpha D - \sigma^{\mu\nu\beta}{}_\alpha F_{\mu\nu}\theta_\beta + \theta^2\sigma^\mu_{\alpha\dot\beta}\partial_\mu\lambda^{*\dot\beta}. \tag{9.24}$$

The gauge transformation of a chiral field of charge q is:

$$\Phi \to e^{-iq\Lambda}\Phi. \tag{9.25}$$

One can form gauge-invariant combinations using the vector field (connection) V:

$$\Phi^\dagger e^{+qV}\Phi. \tag{9.26}$$

We can also define a *gauge-covariant derivative* by

$$\mathcal{D}_\alpha\Phi = D_\alpha\Phi + D_\alpha V\Phi. \tag{9.27}$$

This construction has a non-Abelian generalization. It is most easily motivated by generalizing first the transformation of Φ:

$$\Phi \to e^{-i\Lambda}\Phi \tag{9.28}$$

where Λ is now a matrix-valued chiral field.

Now we want to combine $\phi^\dagger$ and ϕ in a gauge-invariant way. By analogy to what we did in the Abelian case, we introduce a matrix-valued field, V, and require that

$$\Phi^\dagger e^V\Phi \tag{9.29}$$

be gauge-invariant. So we require:

$$e^V \to e^{-i\Lambda^*}e^V e^{i\Lambda}. \tag{9.30}$$

From this, we can define a gauge-covariant field strength,

$$W_\alpha = -\frac{1}{4}\bar D^2 e^{-V}D_\alpha e^V. \tag{9.31}$$

This transforms under gauge transformations like a chiral field in the adjoint representation:

$$W_\alpha \to e^{i\Lambda}W_\alpha e^{-i\Lambda}. \tag{9.32}$$

9.3 $N = 1$ Lagrangians

In ordinary field theories, we construct Lagrangians invariant under translations by integrating densities over all of space. The Lagrangian changes by a derivative under translations, so the *action* is invariant. Similarly, if we start with a Lagrangian density in superspace, a supersymmetry transformation acts by differentiation with respect to x or θ. So integrating the variation over the full superspace

gives zero. This is the basic feature of the integration rules we introduced earlier. In equations:

$$\delta \int d^4x \int d^4\theta\ h(\Phi, \Phi^\dagger, V) = \int d^4x d^4\theta\ (\epsilon^\alpha Q_\alpha + \epsilon_{\dot\alpha} Q^{\dot\alpha}) h(\Phi, \Phi^\dagger, V) = 0. \tag{9.33}$$

For chiral fields, integrals over *half* of superspace are invariant. If $f(\Phi)$ is a function of chiral fields only, f itself is chiral. As a result,

$$\delta \int d^4x\ d^2\theta f(\Phi) = \int d^4x\ d^2\theta (\epsilon^\alpha Q_\alpha + \epsilon_{\dot\alpha} Q^{\dot\alpha}) f(\Phi). \tag{9.34}$$

The integrals over the Q_α terms vanish when integrated over x and $d^2\theta$. The Q^* terms also give zero. To see this, note that $f(\Phi)$ is itself chiral (check), so

$$\bar Q_{\dot\alpha} f \propto \theta^\alpha \sigma^\mu{}_{\alpha\dot\alpha} \partial_\mu f. \tag{9.35}$$

We can construct a general Lagrangian for a set of chiral fields, Φ_i, and gauge group $\mathcal{G}$. The chiral fields have dimension one (again, note that the θs have dimension $-1/2$). The vector fields, V, are dimensionless, while W_α has dimension $3/2$. With these ingredients, we can write down the most general renormalizable Lagrangian. First, there are terms involving integration over the full superspace:

$$\mathcal{L}_{\text{kin}} = \int d^4\theta \sum_i \Phi_i^\dagger e^V \Phi_i, \tag{9.36}$$

where the e^V is in the representation of the gauge group appropriate to the field Φ_i. We can also write an integral over half of superspace:

$$\mathcal{L}_W = \int d^2\theta\, W(\Phi_i) + \text{c.c.} \tag{9.37}$$

$W(\Phi)$ is a holomorphic function of the Φ_is (it is a function of Φ_i, not $\Phi_i^\dagger$), called the superpotential. For a renormalizable theory,

$$W = \frac{1}{2} m_{ij} \Phi_i \Phi_j + \frac{1}{3} \Gamma_{ijk} \Phi^i \Phi_j \Phi_k. \tag{9.38}$$

Finally, for the gauge fields, we can write:

$$\mathcal{L}_{\text{gauge}} = \frac{1}{g^{(i)2}} \int d^2\theta\, W_\alpha^{(i)2}. \tag{9.39}$$

The full Lagrangian density is

$$\mathcal{L} = \mathcal{L}_{\text{kin}} + \mathcal{L}_W + \mathcal{L}_{\text{gauge}}. \tag{9.40}$$

The superspace formulation has provided us with a remarkably simple way to write the general Lagrangian. In this form, however, the meaning of these various

terms is rather opaque. We would like to express them in terms of component fields. We can do this by using our expressions for the fields in terms of their components, and our simple integration table. Let's first consider a single chiral field, Φ, neutral under any gauge symmetries. Then

$$\mathcal{L}_{\text{kin}} = |\partial_\mu \Phi|^2 + i\psi_\Phi \partial_\mu \sigma^\mu \psi_\Phi^* + F_\Phi^* F_\Phi. \tag{9.41}$$

The field F is referred to as an "auxiliary field," as it appears without derivatives in the action. Its equation of motion will be algebraic, and can be easily solved. It has no dynamics. For several fields, labeled with an index i, the generalization is immediate:

$$\mathcal{L}_{\text{kin}} = |\partial_\mu \phi_i|^2 + i\psi_i \partial_\mu \sigma^\mu \psi_i^* + F_i^* F_i. \tag{9.42}$$

It is also easy to work out the component form of the superpotential terms. We will write this for several fields:

$$\mathcal{L}_W = \frac{\partial W}{\partial \Phi_i} F_i + \frac{\partial^2 W}{\partial \Phi_i \Phi_j} \psi_i \psi_j. \tag{9.43}$$

For our special choice of superpotential this is:

$$\mathcal{L}_W = F_i(m_{ij}\Phi_j + \lambda_{ijk}\Phi_j\Phi_k) + (m_{ij} + \lambda_{ijk}\Phi_k)\psi_i\psi_j + \text{c.c.} \tag{9.44}$$

It is a simple matter to solve for the auxiliary fields:

$$F_i^* = -\frac{\partial W}{\partial \Phi_i}. \tag{9.45}$$

Substituting back in the Lagrangian,

$$V = |F_i|^2 = \left| \frac{\partial W}{\partial \Phi_i} \right|^2. \tag{9.46}$$

To work out the couplings of the gauge fields, it is convenient to choose the Wess–Zumino gauge. Again, this is analogous to the Coulomb gauge, in that it makes manifest the physical degrees of freedom (the gauge bosons and gauginos), but the supersymmetry is not explicit. We will leave performing the integrations over superspace to the exercises, and just quote the full Lagrangian in terms of the component fields:

$$\begin{aligned}\mathcal{L} = &-\frac{1}{4} g_a^{-2} F_{\mu\nu}^{a2} - i\lambda^a \sigma^\mu D_\mu \lambda^{a*} + |D_\mu \phi_i|^2 - i\psi_i \sigma^\mu D_\mu \psi_i^* \\ &+ \frac{1}{2g^2}(D^a)^2 + D^a \sum_i \phi_i^* T^a \phi_i + F_i^* F_i - F_i \frac{\partial W}{\partial \phi_i} + \text{c.c.} \\ &+ \sum_{ij} \frac{1}{2} \frac{\partial^2 W}{\partial \phi_i \partial \phi_j} \psi_i \psi_j + i\sqrt{2} \sum \lambda^a \psi_i T^a \phi_i^*. \end{aligned} \tag{9.47}$$

The scalar potential is found by solving for the auxiliary D and F fields:

$$V = |F_i|^2 + \frac{1}{2g_a^2}(D^a)^2 \tag{9.48}$$

with

$$F_i = \frac{\partial W}{\partial \phi_i^*} \quad D^a = \sum_i (g^a \phi_i^* T^a \phi_i). \tag{9.49}$$

In the case there is a $U(1)$ factor in the gauge group, there is one more term one can include in the Lagrangian, known as the Fayet–Iliopoulos D term. In superspace,

$$\xi \int d^4\theta V \tag{9.50}$$

is supersymmetric and gauge invariant, since the integral $\int d^4\theta \Phi$ vanishes for any chiral field. In components, this is simply a term linear in D, ξD, so, solving for D from its equations of motion,

$$D = \xi + \sum_i q_i \phi_i^* \phi_i. \tag{9.51}$$

9.4 The supersymmetry currents

We have written classical expressions for the supersymmetry generators, but for many purposes it is valuable to have expressions for these objects as operators in quantum field theory. We can obtain these by using the Noether procedure. We need to be careful, though, because the Lagrangian is not invariant under supersymmetry transformations, but instead transforms by a total derivative. This is similar to the problem of translations in field theory. To see that there is a total derivative in the variation, recall that the Lagrangian has the form, in superspace:

$$\int d^4\theta f(\theta, \bar{\theta}) + \int d^2\theta W(\theta) + \text{c.c.} \tag{9.52}$$

The supersymmetry generators all involve a $\partial/\partial\theta$ piece and a $\theta\partial_\mu$ piece. The variation of the Lagrangian is proportional to $\int d^4\theta \epsilon Q f + \cdots$. The piece involving $\partial/\partial\theta$ integrates to zero, but the extra piece does not; only in the action, obtained by integrating the Lagrangian density over space-time, does the derivative term drop out.

So in performing the Noether procedure, the variation of the Lagrangian will have the form:

$$\delta\mathcal{L} = \epsilon \partial_\mu K^\mu + (\partial_\mu \epsilon) T^\mu. \tag{9.53}$$

Integrating by parts, we have that $K^\mu - T^\mu$ is conserved. Taking this into account, for a theory with a single chiral field:

$$j^\mu_\alpha = \sqrt{2}\sigma^\nu_{\alpha\dot\beta}\bar\sigma^{\mu\dot\beta\gamma}\psi_\gamma\partial_\nu\phi^* + i\sqrt{2}F\sigma^{\mu\alpha\dot\alpha}\psi^*_{\dot\alpha}, \tag{9.54}$$

and similarly for $j^\mu_{\dot\alpha}$. The generalization for several chiral fields is obvious; one replaces $\psi \to \psi_i$, $\phi \to \phi_i$, etc., and sums over i. One can check that the (anti)commutator of the Qs (integrals over j^0) with the various fields gives the correct transformations laws. One can do the same for gauge fields. Working with the action written in terms of W, there are no derivatives, so the variation of the Lagrangian comes entirely from the $\partial_\mu K^\mu$ term in Eq. (9.53). We have already seen that the variation of $\int d^2\theta$ is a total derivative. The current is worked out in the exercises at the end of this chapter.

9.5 The ground-state energy in globally supersymmetric theories

One striking feature of the Lagrangian of Eq. (9.47) is that $V \geq 0$. This fact can be traced back to the supersymmetry algebra. Start with the equation

$$\{Q_\alpha, \bar Q_{\dot\beta}\} = 2P_\mu\sigma^\mu_{\alpha\dot\beta}, \tag{9.55}$$

multiply by σ^0 and take the trace:

$$Q_\alpha\bar Q_{\dot\alpha} + \bar Q_{\dot\alpha}Q_\alpha = E. \tag{9.56}$$

Since the left-hand side is positive, the energy is always greater than or equal to zero.

In global supersymmetry, $E = 0$ is very special: the expectation value of the energy is an *order parameter* for supersymmetry breaking. If supersymmetry is unbroken, $Q_\alpha|0\rangle = 0$, so the ground-state energy vanishes *if and only if* supersymmetry is unbroken.

Alternatively, consider the supersymmetry transformation laws for λ and ψ. One has, under a supersymmetry transformation with parameter ϵ,

$$\delta\psi = \sqrt{2}\epsilon F + \cdots \quad \delta\lambda = i\epsilon D + \cdots . \tag{9.57}$$

In the quantum theory, the supersymmetry transformation laws become operator equations

$$\delta\psi = i\{Q, \psi\}, \tag{9.58}$$

so taking the vacuum expectation value of both sides, we see that a non-vanishing F means broken supersymmetry; again vanishing of the energy, or not, is an indicator of supersymmetry breaking. So if either F or D has an expectation value, supersymmetry is broken.

The signal of ordinary (bosonic) symmetry breakdown is a Goldstone boson. In the case of supersymmetry, the signal is the presence of a Goldstone fermion, or goldstino. One can prove a goldstino theorem in almost the same way one proves Goldstone's theorem. We will do this shortly, when we consider simple models of supersymmetry and its breaking.

9.6 Some simple models

In this section, we consider some simple models, to develop some practice with supersymmetric Lagrangians and to illustrate how supersymmetry is realized in the spectrum.

9.6.1 The Wess–Zumino model

One of the earliest, and simplest, models is the Wess–Zumino model, a theory of a single chiral field (no gauge interactions). For the superpotential, we take:

$$W = \frac{1}{2}m\phi^2 + \frac{\lambda}{3}\phi^3. \tag{9.59}$$

The scalar potential is (using ϕ for the super and scalar field)

$$V = |m\phi + \lambda\phi^2|^2 \tag{9.60}$$

and the ϕ field has mass-squared $|m|^2$. The fermion mass term is

$$\frac{1}{2}m\psi\psi \tag{9.61}$$

so the fermion also has mass m.

Let's consider the symmetries of the model. First, set $m = 0$. The theory then has a continuous global symmetry. This is perhaps not obvious from the form of the superpotential, $W = (\lambda/3)\phi^3$. But the Lagrangian is an integral over superspace of W, so it is possible for W to transform and for the θs to transform in a compensating fashion. Such a symmetry, which does not commute with supersymmetry, is called an R symmetry. If, by convention, we take the θs to carry charge 1, than the $d\theta$s carry charge -1 (think of the integration rules). So the superpotential must carry charge 2. In the present case, this means that ϕ carries charge $2/3$. Note that each component of the superfield transforms differently:

$$\phi \to e^{i\frac{2}{3}\alpha}\phi \quad \psi \to e^{i(\frac{2}{3}-1)\alpha}\psi \quad F \to e^{i(\frac{2}{3}-2)\alpha}F. \tag{9.62}$$

Now consider the problem of mass renormalization at one loop in this theory. First suppose that $m = 0$. From our experience with non-supersymmetric theories, we might expect a quadratically divergent correction to the scalar mass. But ϕ^2

carries charge 4/3, and this forbids a mass term in the superpotential. For the fermion, the symmetry does not permit one to draw any diagram which corrects the mass. But for the boson, there are two diagrams, one with intermediate scalars, one with fermions. We will study these in detail later. Consistent with our argument, however, these two diagrams cancel.

What if, at tree level, $m \neq 0$? We will see shortly that there are still no corrections to the mass term in the superpotential. In fact, perturbatively, there are no corrections to the superpotential at all. There are, however, wave-function renormalizations; rescaling ϕ corrects the masses. In four dimensions, the wave-function corrections are logarithmically divergent, so there are logarithmically divergent corrections to the masses, but there are no quadratic divergences.

9.6.2 A U(1) gauge theory

Consider a $U(1)$ gauge theory, with two charged chiral fields, ϕ^+ and ϕ^-, with charges ± 1, respectively. First suppose that the superpotential vanishes. Our experience with ordinary field theories would suggest that we start developing a perturbation expansion about the point in field space $\phi^\pm = 0$. But consider the potential in this theory. In Wess–Zumino gauge:

$$V(\phi^\pm) = \frac{1}{2}D^2 = \frac{g^2}{2}(|\phi^+|^2 - |\phi^-|^2)^2. \tag{9.63}$$

Zero energy, supersymmetric minima have $D = 0$. By a gauge choice, we can set

$$\phi^+ = v \quad \phi^- = v' e^{i\alpha}. \tag{9.64}$$

Then $D = 0$ if $v = v'$. In field theory, as discussed in Section 2.3, when one has such a continuous degeneracy, just as in the case of global symmetry breaking, one must choose a vacuum. Each vacuum is physically distinct – in this case, the spectra are different – and there are no transitions between vacua.

It is instructive to work out the spectrum in a vacuum with a given v. One has, first, the gauge bosons, with masses:

$$m_v^2 = 4g^2v^2. \tag{9.65}$$

This accounts for three degrees of freedom. From the Yukawa couplings of the gaugino, λ, to the ϕs, one has a term:

$$\mathcal{L}_\lambda = \sqrt{2}gv\lambda(\psi_{\phi^+} - \psi_{\phi^-}), \tag{9.66}$$

so we have a Dirac fermion with mass $2gv$. So we now have accounted for three bosonic and two fermionic degrees of freedom. The fourth bosonic degree of freedom is a scalar; one can think of it as the "partner" of the Higgs which is eaten in

the Higgs phenomenon. To compute its mass, note that, expanding the scalars as

$$\phi^{\pm} = v + \delta\phi^{\pm} \tag{9.67}$$

$$D = gv(\delta\phi^{+} + \delta\phi^{+*} - \delta\phi^{-} - \delta\phi^{-*}). \tag{9.68}$$

So D^2 gives a mass to the real part of $\delta\phi^{+} - \delta\phi^{-}$, equal to the mass of the gauge bosons and gauginos. Since the masses differ in states with different v, these states are physically inequivalent.

There is also a massless state: a single chiral field. For the scalars, this follows on physical grounds: the expectation value, v, is undetermined and one phase is undetermined, so there is a massless complex scalar. For the fermions, the linear combination $\psi_{\phi^+} + \psi_{\phi^-}$ is massless. So we have the correct number of fields to construct a massless chiral multiplet. We can describe this elegantly by introducing the composite chiral superfield:

$$\Phi = \phi^{+}\phi^{-} \approx v^2 + v(\delta\phi^{+} + \delta\phi^{-}). \tag{9.69}$$

Its components are precisely the massless complex scalar and chiral fermion which we identified above.

This is our first encounter with a phenomenon which is nearly ubiquitous in supersymmetric field theories and string theory. There are often continuous sets of vacuum states, at least in some approximation. The set of such physically distinct vacua is known as the "moduli space." In this example, the set of such states is parameterized by the values of the field, Φ; Φ is called a "modulus."

In quantum mechanics, in such a situation, we would solve for the wave function of the modulus, and the ground state would typically involve a superposition of the different classical ground states. We have seen, though, that, in field theory, one must choose a value of the modulus field. In the presence of such a degeneracy, for each such value one has, in effect, a different theory – no physical process leads to transitions between one such state and another. Once the degeneracy is lifted, however, this is no longer the case, and transitions, as we will frequently see, are possible.

9.7 Non-renormalization theorems

In ordinary field theories, as we integrate out physics between one scale and another, we generate every term in the effective action permitted by symmetries. This is not the case in supersymmetric field theories. This feature gives such theories surprising, and possibly important, properties when we consider questions of naturalness. It also gives us a powerful tool to explore the dynamics of these theories, even at strong coupling. This power comes easily; in this section, we will enumerate these theorems and explain how they arise.

So far, we have restricted our attention to renormalizable field theories. But we have seen that, in considering physics beyond the Standard Model, we may wish to relax this restriction. It is not hard to write down the most general, globally supersymmetric theory with at most two derivatives, using the superspace formalism:

$$\mathcal{L} = \int d^4\theta K(\phi_i, \phi_i^\dagger) + \int d^2\theta W(\phi_i) + \text{c.c.} + \int d^2\theta f_a(\phi)\left(W_\alpha^{(a)}\right)^2 + \text{c.c.} \quad (9.70)$$

The function K is known as the Kahler potential. Its derivatives dictate the form of the kinetic terms for the different fields. The functions W and f_a are holomorphic (what physicists would comfortably call "analytic") functions of the chiral fields. In terms of component fields (see the exercises) the real part of f couples to $F_{\mu\nu}^2$; these functions thus determine the gauge couplings. The imaginary parts couple to the now-familiar operator $F\tilde{F}$. These features of the Lagrangian will be important in much of our discussion of supersymmetric field theories and string theory.

Non-supersymmetric theories have the property that they tend to be generic; any term permitted by symmetries in the theory will appear in the effective action, with an order of magnitude determined by dimensional analysis.[1] Supersymmetric theories are special in that this is not the case. In $N = 1$ theories, there are non-renormalization theorems governing the superpotential and the gauge coupling functions, f, of Eq. (9.70). These theorems assert that the superpotential is not corrected in perturbation theory beyond its tree level value, while f is at most renormalized at one loop.[2]

Originally, these theorems were proven by detailed study of Feynman diagrams. Seiberg has pointed out that they can be understood in a much simpler way. Both the superpotential and the functions f are holomorphic functions of the chiral fields, i.e. they are functions of the ϕ_is and not the ϕ_i^*s. This is evident from their construction. Seiberg argued that the coupling constants of a theory may be thought of as *expectation values* of chiral fields and so the superpotential must be a holomorphic function of these as well. For example, consider a theory of a single chiral field, Φ, with superpotential

$$W = (m\Phi^2 + \lambda\Phi^3). \quad (9.71)$$

We can think of λ and m as expectation values of chiral fields, $\lambda(x, \theta)$ and $m(x, \theta)$.

In the Wess–Zumino Lagrangian, if we first set λ to zero, there is an R symmetry under which Φ has R charge 1 and λ has R charge -1. Now consider corrections to the effective action in perturbation theory. For example, renormalizations of λ in

[1] In some cases, there may be suppression by a few powers of coupling.

[2] There is an important subtlety connected with these theorems. Both should be interpreted as applying only to a "Wilsonian" effective action, in which one integrates out physics above some scale, μ. If infrared physics is included, the theorems do not necessarily hold. This is particularly important for the gauge couplings.

the superpotential necessarily involve positive powers of λ. But such terms (apart from $(\lambda)^1$) have the wrong R charge to preserve the symmetry. So there can be no renormalization of this coupling. There can be wave-function renormalization, since K is not holomorphic, so $K = K(\lambda^\dagger\lambda)$ is allowed, in general.

There are many interesting generalizations of these ideas, and I won't survey them here, but I will mention two further examples. First, gauge couplings can be thought of in the same way, i.e. we can treat g^{-2} as part of a chiral field. More precisely, we define:

$$S = \frac{8\pi^2}{g^2} + ia + \cdots . \tag{9.72}$$

The real part of the scalar field in this multiplet couples to $F_{\mu\nu}^2$, but the imaginary part, a, couples to $F\tilde{F}$. Because $F\tilde{F}$ is a total derivative, in perturbation theory there is a symmetry under constant shifts of a. The effective action should respect this symmetry. Because the gauge coupling function, f, is holomorphic, this implies that

$$f(g^2) = S + \text{const} = \frac{8\pi^2}{g^2} + \text{const.} \tag{9.73}$$

The first term is just the tree level term. One-loop corrections yield a constant, but there are no higher order corrections in perturbation theory! This is quite a striking result. It is also paradoxical, since the two-loop beta functions for supersymmetric Yang–Mills theories have been computed long ago, and are, in general, non-zero. The resolution of this paradox is subtle and interesting. It provides a simple computation of the two-loop beta function. In a particular renormalization scheme, it gives an exact expression for the beta function. This is explained in Appendix D.

Before explaining the resolution of this paradox, there is one more non-renormalization theorem which we can prove rather trivially here. This is the statement that if there is no Fayet–Iliopoulos D term at tree level, this term can be generated at most at one loop. To prove this, write the D term as

$$\int d^4\theta d(g, \lambda)V. \tag{9.74}$$

Here $d(g, \lambda)$ is some unknown function of the gauge and other couplings in the theory. But if we think of g and λ as chiral fields, this expression is only gauge invariant if d is a constant, corresponding to a possible one-loop contribution. Such contributions do arise in string theory.

In string theory, all of the parameters *are* expectation values of chiral fields. Indeed, non-renormalization theorems in string theory, both for world sheet and string perturbation theory, were proved by the sort of reasoning we have used above.

9.8 Local supersymmetry: supergravity

If supersymmetry has anything to do with nature, and if it is not merely an accident, then it must be a local symmetry. There is not space here for a detailed exposition of local supersymmetry. For most purposes, both theoretical and phenomenological, there are, fortunately, only a few facts we will need to know. The field content (in four dimensions) is like that of global supersymmetry, except that now one has a graviton and a gravitino. Note that the number of additional bosonic and fermionic degrees of freedom (a minimal requirement if the theory is to be supersymmetric) is the same. The graviton is a traceless, symmetric tensor; in $d-2=2$ dimensions, this has two independent components. Similarly, the gravitino, ψ_μ, has both a vector and a spinor index. It satisfies a constraint similar to tracelessness:

$$\gamma^\mu \psi_\mu = 0. \tag{9.75}$$

In $d-2$ dimensions, this is two conditions, leaving two physical degrees of freedom.

As in global supersymmetry (without the restriction of renormalizability), the terms in the effective action with at most two derivatives or four fermions are completely specified by three functions.

(1) The Kahler potential, $K(\phi, \phi^\dagger)$, a function of the chiral fields.
(2) The superpotential, $W(\phi)$, a holomorphic function of the chiral fields.
(3) The gauge coupling functions, $f_{\rm a}(\phi)$, which are also holomorphic functions of the chiral fields.

The Lagrangian which follows from these is quite complicated, including many two- and four-fermion interactions. It can be found in the suggested reading. Our main concern in this text will be the scalar potential. This is given by

$$V = e^K \left[\left(\frac{\partial W}{\partial \phi_i} + \frac{\partial K}{\partial \phi_i} W \right) g^{i\bar{j}} \left(\frac{\partial W^*}{\partial \phi_{\bar{j}}^*} + \frac{\partial K}{\partial \phi_{\bar{j}}^*} W \right) - 3|W|^2 \right], \tag{9.76}$$

where

$$g_{i\bar{j}} = \frac{\partial^2 K}{\partial \phi_i \partial \phi_{\bar{j}}} \tag{9.77}$$

is the (Kahler) metric associated with the Kahler potential. In this equation, we have adopted units in which $M=1$, where

$$G_N = \frac{1}{8\pi M^2}. \tag{9.78}$$

$M \approx 2 \times 10^{18}$ GeV is known as the *reduced Planck mass*.

Suggested reading

The text by Wess and Bagger (1992) provides a good introduction to superspace, the fields and Lagrangians of supersymmetric theories in four dimensions, and supergravity. Other texts include those by Gates *et al*. (1983) and Mohapatra (2003). Appendix B of Polchinski's (1998) text provides a concise introduction to supersymmetry in higher dimensions. The supergravity Lagrangian is derived and presented in its entirety in Cremmer *et al*. (1979) and Wess and Bagger (1992) and is reviewed, for example, in Nilles (1984). Non-renormalization theorems were first discussed from the viewpoint presented here by Seiberg (1993).

Exercises

(1) Verify the commutators of the Qs and the Ds.
(2) Check that with the definition, Eq. (9.15), Φ is chiral. Show that any function of chiral fields is a chiral field.
(3) Verify that W_α transforms as in Eq. (9.32), and that $\mathrm{Tr} W_\alpha^2$ is gauge invariant.
(4) Derive the expression (9.47) for the component Lagrangian including gauge interactions and the superpotential by doing the superspace integrals. For an $SU(2)$ theory with a scalar triplet, $\vec{\phi}$, and singlet, X, take $W = \lambda(\vec{\phi}^2 - \mu^2)$. Find the ground state and work out the spectrum.
(5) Derive the supersymmetry current for a theory with several chiral fields. For a single field, Φ, and $W = 1/2\, m\Phi^2$, verify, using the canonical commutation relations, that the Qs obey the supersymmetry algebra. Work out the supercurrent for a pure supersymmetric gauge theory.

10

A first look at supersymmetry breaking

If supersymmetry has anything to do with the real world, it must be a broken symmetry. In the globally supersymmetric framework we have presented so far, this breaking could be spontaneous or explicit. Once we promote the symmetry to a local symmetry, as we will argue later, the breaking of supersymmetry must be spontaneous. However, as we will also see, at low energies, the theory can appear to be a globally supersymmetric theory with explicit, "soft," breaking of the symmetry. In this chapter, we will discuss some of the features of both spontaneous and explicit breaking.

10.1 Spontaneous supersymmetry breaking

We have seen that supersymmetry breaking is signalled by a non-zero expectation value of an F component of a chiral or D component of a vector superfield. Models involving only chiral fields with no supersymmetric ground state are referred to as O'Raifeartaigh models. A simple example has three singlet fields, A, B, and X, with superpotential:

$$W = \lambda A(X^2 - \mu^2) + mBX. \tag{10.1}$$

With this superpotential, the equations

$$F_A = \frac{\partial W}{\partial A} = 0 \quad F_B = \frac{\partial W}{\partial B} = 0 \tag{10.2}$$

are incompatible. To actually determine the expectation values and the vacuum energy, it is necessary to minimize the potential. There is no problem satisfying the equation $F_X = 0$. So we need to minimize

$$V_{\text{eff}} = |F_A|^2 + |F_B|^2 = |\lambda^2||X^2 - \mu^2|^2 + m^2|X|^2. \tag{10.3}$$

Assuming μ^2 and λ are real, the solutions are:

$$X = 0 \quad X^2 = \frac{2\lambda^2\mu^2 - m^2}{2\lambda^2}. \tag{10.4}$$

The corresponding vacuum energies are:

$$V_0^{(a)} = |\lambda^2\mu^4|; \quad V_0^{(b)} = m^2\mu^2 - \frac{m^4}{4\lambda^2}. \tag{10.5}$$

The vacuum at $X \neq 0$ disappears at a critical value of μ.

Let's consider the spectrum in the first of these (with $X = 0$). We focus, in particular, on the massless states. First, there is a massless scalar. This arises because, at this level, not all of the fields are fully determined. The equation

$$\frac{\partial W}{\partial X} = 0 \tag{10.6}$$

can be satisfied provided

$$2\lambda AX + mB = 0. \tag{10.7}$$

This vacuum degeneracy is accidental, and as we will later see, is lifted by quantum corrections.

There is also a massless fermion, ψ_A. This fermion is the goldstino. Replacing the auxiliary fields in the supersymmetry current for this model (Eq. (9.54)) gives

$$j_\mu^\alpha = i\sqrt{2}F_A\sigma^\mu_{\alpha\dot{\alpha}}\psi_A^{*\dot{\alpha}}. \tag{10.8}$$

You should check that the massive states do not form Bose–Fermi degenerate multiplets.

10.1.1 The Fayet–Iliopoulos D term

It is also possible to generate an expectation value for a D term. In the case of a $U(1)$ gauge symmetry, we saw that

$$\mu^2 \int d^4\theta\ V = \mu^2 D \tag{10.9}$$

is gauge invariant. Under the transformation $\delta V = \Lambda + \Lambda^\dagger$, the integrals over the chiral and anti-chiral fields, Λ and $\Lambda^\dagger$, are zero. This can be seen either by doing the integration directly, or by noting that differentiation by Grassmann numbers is equivalent to integration (recall our integral table). As a result, for example, $\int d^2\bar{\theta} \propto (\bar{D})^2$. This Fayet–Iliopoulos D term can lead to supersymmetry breaking. For example, if one has two charged fields, $\Phi^\pm$, with charges ± 1, and superpotential $m\Phi^+\Phi^-$, one cannot simultaneously make the two auxiliary F fields and the auxiliary D field vanish.

One important feature of both types of models is that at tree level, in the context of global supersymmetry, the spectra are never realistic. These spectra satisfy a sum rule,

$$\sum (-1)^F m^2 = 0. \tag{10.10}$$

Here $(-1)^F = 1$ for bosons and -1 for fermions. This guarantees that there are always light states, and often color and/or electromagnetism are broken. These statements are not true of radiative corrections, and of supergravity, as we will explain later.

It is instructive to prove this sum rule. Consider a theory of chiral fields only (no gauge interactions). The potential is given by

$$V = \sum_i \left| \frac{\partial W}{\partial \phi_i} \right|^2 . \tag{10.11}$$

The boson mass matrix has terms of the form $\phi_{\bar{i}}^* \phi_j$ and $\phi_i \phi_j$ + c.c., where we are using indices $\bar{i}$ and $\bar{j}$ for complex conjugate fields. The latter terms, as we will now see, are connected with supersymmetry breaking. The various terms in the mass matrix can be obtained by differentiating the potential:

$$m^2_{i\bar{j}} = \frac{\partial^2 V}{\partial \phi_i \partial \phi^*_{\bar{j}}} = \frac{\partial^2 W}{\partial \phi_i \partial \phi_k} \frac{\partial^2 W^*}{\partial \phi^*_{\bar{k}} \partial \phi^*_{\bar{j}}}, \tag{10.12}$$

$$m^2_{ij} = \frac{\partial^2 V}{\partial \phi_i \partial \phi_j} = \frac{\partial W}{\partial \phi^*_{\bar{k}}} \frac{\partial^3 W}{\partial \phi_k \partial \phi_i \partial \phi_j}. \tag{10.13}$$

The first of these terms has just the structure of the square of the fermion mass matrix,

$$\mathcal{M}_{\mathrm{F}ij} = \frac{\partial^2 W}{\partial \phi_i \partial \phi_j}. \tag{10.14}$$

So writing the boson mass, $\mathcal{M}^2_{\mathrm{B}}$ matrix on the basis $(\phi_i \ \phi^*_{\bar{j}})$, we see that Eq. (10.10) holds.

The theorem is true whenever a theory can be described by a renormalizable effective action. Various non-renormalizable terms in the effective action can give additional contributions to the mass. For example, in our O'Raifeartaigh model, $\int d^4\theta A^\dagger A Z^\dagger Z$ will violate the tree-level sum rule. Such terms arise in renormalizable theories when one integrates out heavy fields to obtain an effective action at some scale. In the context of supergravity, such terms are present already at tree level. This is perhaps not surprising, given that these theories are non-renormalizable and must be viewed as effective theories from the very beginning (perhaps the effective low-energy description of string theory). Shortly, we will

discuss the construction of realistic models. First, however, we turn to the issue of non-renormalization theorems and dynamical supersymmetry breaking.

10.2 The goldstino theorem

In each of the examples of supersymmetry breaking, there is a massless fermion in the spectrum. We might expect this, by analogy to Goldstone's theorem. The essence of the usual Goldstone theorem is the statement that, for a spontaneously broken global symmetry, there is a massless scalar. There is a coupling of this scalar to the symmetry current. From Lorentz invariance (see Appendix B):

$$\langle 0|j^\mu|\pi(p)\rangle = fp^\mu. \tag{10.15}$$

Correspondingly, in the low-energy effective field theory (valid below the scale of symmetry breaking) the current takes the form:

$$j^\mu = f\partial^\mu\pi(x). \tag{10.16}$$

The analogous statements for spontaneous breaking of global supersymmetry are easy to prove. Suppose that the symmetry is broken by the F component of a chiral field (this can be a composite field). Then we can study

$$\int d^4x\,\partial_\mu e^{iq\cdot x}T\left\langle j^\mu_\alpha(x)\psi_\Phi(0)\right\rangle = 0. \tag{10.17}$$

Here j^μ_α is the supersymmetry current; its integral over space is the supersymmetry charge. This expression vanishes because it is an integral of a total derivative. Now taking the derivatives, there are two non-vanishing contributions: one from the derivative acting on the exponential; one from the action on the time-ordering symbol. Taking these derivatives, and then taking the limit $q \to 0$, gives

$$\langle\{Q, \Psi_\Phi(0)\}\rangle = iq_\mu T\left\langle j^\mu_\alpha(x)\psi_\Phi(0)\right\rangle_{\text{F.T.}}. \tag{10.18}$$

The left-hand side is constant, so the Green function on the right-hand side must be singular as $q \to 0$. By the usual spectral representation analysis, this shows that there is a massless fermion coupled to the supersymmetry current. In weakly coupled theories, we can understand this more simply. Recalling the form of the supersymmetry current, if one of the Fs has an expectation value,

$$j^\mu_\alpha = i\sqrt{2}(\sigma^\mu)_{\alpha\dot\alpha}\psi^{*\dot\alpha}F. \tag{10.19}$$

To leading order in fields, current conservation is just the massless Dirac equation. F, here, is the "goldstino decay constant." We can understand the massless fermion which appeared in the O'Raifeartaigh model in terms of this theorem. It is easy to check that:

$$\psi_G \propto F_A\psi_A + F_B\psi_B. \tag{10.20}$$

10.3 Loop corrections and the vacuum degeneracy

We saw that in the O'Raifeartaigh model, at the classical level, there is a large vacuum degeneracy. To understand the model fully, we need to investigate the fate of this degeneracy in the quantum theory. Consider the vacuum with $X = 0$. In this case, A is undetermined at the classical level. But A is only an approximate modulus. At one loop, quantum corrections generate a potential for A. Our goal is to integrate out the various massive fields to obtain the effective action for A. At one loop, this is particularly easy. The tree-level mass spectrum depends on A. The one-loop vacuum energy is:

$$\sum_i (-1)^F \int \frac{d^3k}{(2\pi)^3} \frac{1}{2} \sqrt{\vec{k}^2 + m_i^2}. \tag{10.21}$$

Here the sum is over all possible helicity states; again the factor $(-1)^F$ weights bosons with 1 and fermions with -1. This expression, in field theory, is usually very divergent in the ultraviolet, but in the supersymmetric case, it is far less so. If supersymmetry is unbroken, the boson and fermion contributions cancel, and the correction simply vanishes. If supersymmetry is broken, the divergence is only logarithmic. To see this, we can simply study the integrand at large k, expanding the square root in powers of m^2/k^2. The leading, quartically divergent term, is independent of m^2, and so vanishes. The next term is quadratically divergent, but it vanishes because of the sum rule: $\sum (-1)^F m_i^2 = 0$.

So at one loop the potential behaves as:

$$V(A) = -\sum (-1)^F m_i^4 \int \frac{d^3k}{16(2\pi)^3 k^3} \approx \sum (-1)^F m_i^4 \frac{1}{64\pi^2} \ln\left(m_i^2/\Lambda^2\right). \tag{10.22}$$

To compute the potential precisely, we need to work out the spectrum as a function of A. We will content ourselves with the limit of large A. Then the spectrum consists of a massive fermion, ψ_X, with mass $2\lambda A$, and the real and imaginary parts of the scalar components of X, with masses:

$$m_s^2 = 4|\lambda^2 A^2| \pm 2\mu^2 \lambda^2 x^2. \tag{10.23}$$

So

$$V(A) = |\lambda^2|\mu^4 \left(1 + \frac{\lambda^2}{8\pi^2} \ln(|\lambda A|^2/\Lambda^2)\right). \tag{10.24}$$

This result has a simple interpretation. The leading term is the classical energy; the correction corresponds to replacing λ^2 by $\lambda^2(A)$, the running coupling at scale A. In this theory, a more careful study shows that the minimum of the potential is precisely at $A = 0$.

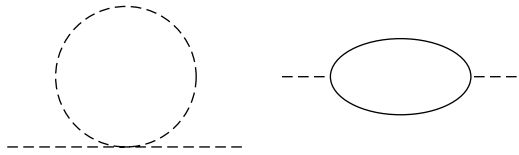

Fig. 10.1. One-loop corrections to scalar masses arising from Yukawa couplings.

10.4 Explicit, soft supersymmetry breaking

Ultimately, if nature is supersymmetric, it is likely that we will want to understand supersymmetry breaking through some dynamical mechanism. But we can be more pragmatic, accept that supersymmetry is broken, and parameterize the breaking through mass differences between ordinary fields and their superpartners. It turns out that this procedure does not spoil the good ultraviolet properties of the theory. Such mass terms are said to be "soft" for precisely this reason.

We will consider soft breakings in more detail in the next chapter when we discuss the MSSM, but we can illustrate the main point simply. Take as a model the Wess–Zumino model, with $m = 0$ in the superpotential. Add to the Lagrangian an explicit mass term $m_{\rm soft}^2|\phi|^2$. Then we can calculate the one-loop correction to the scalar mass from the two graphs of Fig. 10.1. In the supersymmetric case, these two graphs cancel. With the soft breaking term, there is not an exact cancellation; instead one obtains:

$$\delta m^2 = -\frac{|\lambda|^2}{16\pi^2} m_{\rm soft}^2 \ln\left(\Lambda^2/m_{\rm soft}^2\right). \tag{10.25}$$

We can understand this simply on dimensional grounds. We know that, for $m_{\rm soft}^2 = 0$, there is no correction. Treating the soft term as a perturbation, the result is necessarily proportional to $m_{\rm soft}^2$; at most, then, any divergence must be logarithmic.

In addition to soft masses for scalars, one can also add soft masses for gauginos; one can also include trilinear scalar couplings. We can understand how these might arise at a more fundamental level, which also makes clear the sense in which these terms are soft. Suppose that we have a field, Z, with non-zero F component, as in the O'Raifeartaigh model (but more generally). Suppose at tree level, there are no renormalizable couplings between Z and the other fields of the model, which we will denote generically as ϕ. Non-renormalizable couplings, such as

$$\mathcal{L}_Z = \frac{1}{M^2}\int d^4\theta Z^\dagger Z \phi^\dagger \phi, \tag{10.26}$$

can be expected to arise in the effective Lagrangian; they are not forbidden by any symmetry. Replacing Z by its expectation value, $\langle Z\rangle = \cdots + \theta^2\langle F_Z\rangle$, gives a mass

term for the scalar component of ϕ:

$$\mathcal{L}_Z = \frac{|\langle F \rangle|^2}{M^2} |\phi|^2 + \cdots \tag{10.27}$$

This is precisely the soft scalar mass we described above; it is soft because it is associated with a high-dimension operator. Similarly, the operator:

$$\int d^2\theta \frac{Z}{M} W_\alpha^2 = \frac{F_Z}{M} \lambda\lambda + \cdots \tag{10.28}$$

gives rise to a mass for gauginos. The term

$$\int d^2\theta \frac{Z}{M} \phi\phi\phi \tag{10.29}$$

leads to a trilinear coupling of the scalars. Simple power counting shows that loop corrections to these couplings due to renormalizable interactions are at most logarithmically divergent.

To summarize, there are three types of soft breaking terms which can appear in a low-energy effective action.

- Soft scalar masses, $m_\phi^2 |\phi|^2$ and $\tilde{m}\phi^2\phi\phi + c.c.$
- Gaugino masses, $m_\lambda \lambda\lambda$.
- Trilinear scalar couplings, $\Gamma\phi\phi\phi$.

All three types of couplings will play an important role in thinking about possible supersymmetry phenomenologies.

10.5 Supersymmetry breaking in supergravity models

We stressed in the last chapter that, since nature includes gravity, if supersymmetry is not simply an accident, it must be a local symmetry. If the underlying scale of supersymmetry breaking is high enough, supergravity effects will be important. The potential of a supergravity model will be sufficiently important to us that it is worth writing again:

$$V = e^K \left[\left(\frac{\partial W}{\partial \phi_i} + \frac{\partial K}{\partial \phi_i} W \right) g^{i\bar{j}} \left(\frac{\partial W}{\partial \phi_{\bar{j}}^*} + \frac{\partial K}{\partial \phi_{\bar{j}}^*} W^* \right) - 3|W|^2 \right]. \tag{10.30}$$

In supergravity, the condition for unbroken supersymmetry is that the *Kahler derivative* of the superpotential should vanish:

$$D_i W = \frac{\partial W}{\partial \phi_i} + \frac{\partial K}{\partial \phi_i} W = 0. \tag{10.31}$$

When this is not the case, supersymmetry is broken. If we require vanishing of the cosmological constant, then we have:

$$3|W|^2 = \sum_{i,\bar{j}} D_i W D_{\bar{j}} W^* g^{i\bar{j}}. \tag{10.32}$$

In this case, the gravitino mass turns out to be:

$$m_{3/2} = \left\langle e^{K/2} W \right\rangle. \tag{10.33}$$

There is a standard strategy for building supergravity models. One introduces two sets of fields, the "hidden sector fields," which will be denoted by Z_i, and the "visible sector fields," denoted by y_a. The Z_is are assumed to be connected with supersymmetry breaking, and to have only very small couplings to the ordinary fields, y_a. In other words, one assumes that the superpotential, W, has the form

$$W = W(Z) + W_y(y), \tag{10.34}$$

at least up to terms suppressed by $1/M$. The y fields should be thought of as the ordinary matter fields and their superpartners.

One also usually assumes that the Kahler potential has a "minimal" form,

$$K = \sum Z_i^\dagger Z_i + \sum y_a^\dagger y_a. \tag{10.35}$$

One chooses (tunes) the parameters of W_Z so that

$$\langle F_Z \rangle \approx M_W M \tag{10.36}$$

and

$$\langle V \rangle = 0. \tag{10.37}$$

Note that this means that

$$\langle W \rangle \approx M_W M^2. \tag{10.38}$$

The simplest model of the hidden sector is known as the "Polonyi model." In this model,

$$W = m^2(Z + \beta) \tag{10.39}$$

$$\beta = (2 + \sqrt{3})M. \tag{10.40}$$

In global supersymmetry, with only renormalizable terms, this would be a rather trivial superpotential, but not so in supergravity. The minimum of the potential for Z lies at

$$Z = (\sqrt{3} - 1)M \tag{10.41}$$

and

$$m_{3/2} = (m^2/M)e^{(\sqrt{3}-1)^2/2}. \tag{10.42}$$

This symmetry breaking also leads to soft breaking mass terms for the fields y. There are terms of the form

$$m_0^2|y_i|^2. \tag{10.43}$$

These arise from the $|\partial_i K\ W|^2 = |y_i|^2|W|^2$ terms in the potential. For the simple Kahler potential:

$$m_0^2 = 2\sqrt{3}m_{3/2}^2 \quad A = (3-\sqrt{3})m_{3/2}. \tag{10.44}$$

If we now allow for a non-trivial W_y, we also find supersymmetry-violating quadratic and cubic terms in the potential. These are known as the B and A terms, and have the form:

$$B_{ij}m_{3/2}\phi_i\phi_j + A_{ijk}m_{3/2}\phi_i\phi_j\phi_k. \tag{10.45}$$

For example, if W is homogeneous, and of degree three, there are terms in the supergravity potential of the form:

$$e^K \frac{\partial W}{\partial y_a}\frac{\partial K}{\partial y_a^*}\langle W\rangle + \text{c.c.} = 3m_{3/2}W(y). \tag{10.46}$$

Additional contributions arise from

$$e^K \left\langle \frac{\partial W}{\partial z_i} \right\rangle \langle z_i^*\rangle W^* + \text{c.c.} \tag{10.47}$$

There are analogous contributions to the B terms. In the exercises, these are worked out for specific models.

Gaugino masses (both in local and global supersymmetry) can arise from a non-trivial gauge coupling function,

$$f^a = c\frac{Z}{M}, \tag{10.48}$$

which gives

$$m_\lambda = \frac{cF_z}{M}. \tag{10.49}$$

These models have just the correct structure to build a theory of TeV-scale supersymmetry, provided $m_{3/2} \sim$ TeV. They have soft breakings of the correct order of magnitude. We will discuss their phenomenology further when we discuss the Minimal Supersymmetric Standard Model in the next chapter.

Even without a deep understanding of local supersymmetry, there are a number of interesting observations we can make. Most important, our arguments for the renormalization of the superpotential in global supersymmetry remain valid here. This will be particularly important when we come to string theory, which is a locally supersymmetric theory.

Suggested reading

It was Witten (1981) who most clearly laid out the issues of supersymmetry breaking. This paper remains extremely useful and readable today. The notion that one should consider adding soft breaking parameters to the MSSM was developed by Dimopoulos and Georgi (1981). Good introductions to models with supersymmetry breaking in supergravity are provided by a number of review articles and textbooks, for example those of Mohapatra (2003) and Nilles (1984).

Exercises

(1) Work out out the spectrum of the O'Raifeartaigh model. Show that the spectrum is not supersymmetric, but verify the sum rule, $\sum(-1)^F m^2 = 0$.
(2) Work out the spectrum of a model with a Fayet–Iliopoulos D term and supersymmetry breaking. Again verify the sum rule.
(3) Check Eqs. (10.40)–(10.44) for the Polonyi model.

11

The Minimal Supersymmetric Standard Model

We can now very easily construct a supersymmetric version of the Standard Model. For each of the gauge fields of the usual Standard Model, we introduce a vector superfield. For each of the fermions (quarks and leptons) we introduce a chiral superfield with the same gauge quantum numbers. Finally, we need at least two Higgs doublet chiral fields; if we introduce only one, as in the simplest version of the Standard Model, the resulting theory possesses gauge anomalies and is inconsistent. So the theory is specified by the gauge group $(SU(3) \times SU(2) \times U(1))$ and enumerating the chiral fields:

$$Q_f, \bar{u}_f, \bar{d}_f, L_f, \bar{e}_f \quad f = 1, 2, 3 \quad H_U, H_D. \tag{11.1}$$

The gauge invariant kinetic terms, auxiliary D terms, and gaugino–matter Yukawa couplings are completely specified by the gauge symmetries. The superpotential can be taken to be

$$W = H_U(\Gamma_U)_{f,f'} Q_f \bar{U}_{f'} + H_D(\Gamma_D)_{f,f'} Q_f \bar{D}_{f'} + H_D(\Gamma_E)_{f,f'} L_f \bar{e}_{f'}. \tag{11.2}$$

If the Higgs obtain suitable expectation values, $SU(2) \times U(1)$ is broken, and quarks and leptons acquire mass, just as in the Standard Model.

There are other terms which can also be present in the superpotential. These include the "μ-term," $\mu H_U H_D$. This is a supersymmetric mass term for the Higgs fields. We will see later that we need $\mu \gtrsim M_Z$ to have a viable phenomenology. A set of dimension four terms permitted by the gauge symmetries raise serious issues. For example, one can have terms

$$\bar{u}_f \bar{d}_g \bar{d}_h \Gamma^{fgh} + Q_f L_g \bar{d}_h \lambda^{fgh}. \tag{11.3}$$

These couplings violate B and L! This is our first serious setback. In the Standard Model, there is no such problem. The leading B and L violating operators permitted by gauge invariance possess dimension six, and they will be highly suppressed if

the scale of interactions which violate these symmetries is high, as in grand unified theories.

If we are not going to simply give up, we need to suppress B and L violation at the level of dimension-four terms. The simplest approach is to postulate additional symmetries. There are various possibilities one can imagine.

(1) Global continuous symmetries. It is hard to see how such symmetries could be preserved in any quantum theory of gravity, and in string theory, there is a theorem which asserts that there are no global continuous symmetries.
(2) Discrete symmetries. As we will see later, discrete symmetries can be gauge symmetries. As such, they will not be broken in a consistent quantum theory. They are common in string theory. These symmetries are often "R-symmetries," symmetries which do not commute with supersymmetry.

A simple (though not unique) solution to the problem of baryon and lepton number violation by dimension-four operators is to postulate a discrete symmetry known as R-parity. Under this symmetry, all ordinary particles are even, while their superpartners are odd. Imposing this symmetry immediately eliminates all of the dangerous operators. For example,

$$\int d^2\theta \bar{u}\bar{d}\bar{d} \sim \psi_{\bar{u}}\psi_{\bar{d}}\tilde{\bar{d}} \tag{11.4}$$

(we have changed notation again: the tilde here indicates the superpartner of the ordinary field, i.e. the squark) is odd under the symmetry.

More formally, we can define this symmetry as the transformation on superfields:

$$\theta_\alpha \to -\theta_\alpha \tag{11.5}$$

$$(Q_f, \bar{u}_f, \bar{d}_f, L_f, \bar{e}_f) \to -(Q_f, \bar{u}_f, \bar{d}_f, L_f, \bar{e}_f) \tag{11.6}$$

$$(H_U, H_D) \to (H_U, H_D). \tag{11.7}$$

With this symmetry, the full, renormalizable superpotential is just that of Eq. (11.2).

In addition to solving the problem of very fast proton decay, R-parity has another striking consequence: the lightest of the new particles predicted by supersymmetry (the LSP) is *stable*. This particle can easily be neutral under the gauge groups. It is then, inevitably, very weakly interacting. This in turn means the following.

- The generic signature of R-parity conserving supersymmetric theories is events with missing energy.
- Supersymmetry is likely to produce an interesting dark-matter candidate.

This second point is one of the principal reasons that many physicists have found the possibility of low-energy supersymmetry so compelling. If one calculates the dark-matter density, as we will see in the chapter on cosmology, one

automatically finds a density in the right range if the scale of supersymmetry breaking is about 1 TeV. Later we will see an additional piece of circumstantial evidence for low-energy supersymmetry: the unification of the gauge couplings within the MSSM.

We can imagine more complicated symmetries which would have similar effects, and we will have occasion to discuss these later. In most of what follows, we will assume a conserved Z_2 R-parity.

11.1 Soft supersymmetry breaking in the MSSM

If supersymmetry is a feature of the underlying laws of nature, it is certainly broken. The simplest approach to model building with supersymmetry is to add soft-breaking terms to the effective Lagrangian so that the squarks, sleptons and gauginos have sufficiently large masses that they have not yet been observed (or, in the event that they are discovered, to account for their values). On the other hand, these masses shouldn't be so large that they reintroduce the fine-tuning problem.

Without a microscopic theory of supersymmetry breaking, all of the soft terms are independent. It is interesting to ask, in the MSSM, how many soft-breaking parameters are there? More precisely, let's count the parameters of the model beyond those of the minimal standard model with a single Higgs doublet. Having imposed R-parity, the number of Yukawa couplings is the same in both theories, as are the number of gauge couplings and θ parameters. The quartic couplings of the Higgs fields are completely determined in terms of the gauge couplings. So the "new" terms arise from the soft-breaking terms, as well as the μ term for the Higgs fields. We will speak loosely of all of this as the "soft-breaking" Lagrangian. Suppressing flavor indices:

$$\begin{aligned}\mathcal{L}_{\rm sb} = {} & \widetilde{Q}^* m_Q^2 \widetilde{Q} + \tilde{\bar{u}}^* m_{\bar{u}}^2 \tilde{\bar{u}} + \tilde{\bar{d}}^* m_{\bar{d}}^2 \tilde{\bar{d}} + \tilde{L}^* m_L^2 \tilde{L} + \tilde{\bar{e}}^* m_{\bar{e}}^2 \tilde{\bar{e}} + H_U \widetilde{Q} A_u \tilde{\bar{u}} + H_D \widetilde{Q} A_d \tilde{\bar{d}} \\ & + H_D \widetilde{L} \widetilde{A}_l \tilde{\bar{e}} + \text{c.c.} + M_i \lambda\lambda + \text{c.c.} + m_{H_U}^2 |H_U|^2 + m_{H_D}^2 |H_D|^2 + \mu B H_U H_D \\ & + \mu \psi_H \psi_H . \end{aligned} \tag{11.8}$$

The matrices m_Q^2, $m_{\bar{u}}^2$, and so on are 3×3 Hermitian matrices, so they have 9 independent entries. The matrices A_u, A_d, etc., are general 3×3 complex matrices, so they each possess 18 independent entries. Each of the gaugino masses is a complex number, so these introduce 6 additional parameters. The quantities μ and B are also complex; this is 4 more. In total, then, there are 111 new parameters. As in the standard model, not all of these parameters are meaningful; we are free to make field redefinitions. The counting is significantly simplified if we just ask how many parameters there are beyond the usual 18 of the minimal theory, since this counting uses up most of this freedom.

To understand what redefinitions are possible beyond the transformations on the quarks and leptons which go into defining the KM parameters, we need to ask what are the symmetries of the MSSM before introducing the soft-breaking terms and the μ term (the μ term is more or less on the same footing as the soft-breaking terms, since it is of the same order of magnitude; as we will discuss later, it might well arise from the physics of supersymmetry breaking). Apart from the usual baryon and lepton numbers, there are two more. The first is a Peccei–Quinn symmetry, under which two Higgs superfields rotate by the same phase, while the right-handed quarks and leptons rotate by the opposite phase. The second is an R-symmetry, a generalization of the symmetry we found in the Wess–Zumino model. It is worth describing this in some detail. By definition, an R-symmetry is a symmetry of the Hamiltonian which does not commute with the supersymmetry generators. Such symmetries can be continuous or discrete. In the case of continuous R-symmetries, by convention, we can take the θs to transform by a phase $e^{i\alpha}$. Then the general transformation law takes the form

$$\lambda_i \to e^{i\alpha}\lambda_i \tag{11.9}$$

for the gauginos, while for the elements of a chiral multiplet

$$\Phi_i(x,\theta) \to e^{ir_i\alpha}\Phi(x,\theta e^{i\alpha}), \tag{11.10}$$

or, in terms of the component fields,

$$\phi_i \to e^{ir_i\alpha}\phi_i \quad \psi_i \to e^{i(r_i-1)\alpha}\psi_i \quad F_i \to e^{i(r_i-2)\alpha}F_i. \tag{11.11}$$

In order that the Lagrangian exhibit a continuous R-symmetry, the total R charge of all terms in the superpotential must be 2. In the MSSM, we can take $r_i = 2/3$ for all of the chiral fields.

The soft-breaking terms, in general, break two of the three lepton number symmetries, the R-symmetry and the Peccei–Quinn symmetry. So there are 4 non-trivial field redefinitions which we can perform. In addition, the minimal standard model has 2 Higgs parameters. So from our 111 parameters, we can subtract a total of 6, leaving 105 as the number of *new* parameters in the MSSM.

Clearly we would like to have a theory which predicts these parameters. Later, we will study some candidates. To get started, however, it is helpful to make an ansatz. The simplest thing to do is suppose that all of the scalar masses are the same, all of the gaugino masses the same, and so on. It is necessary to specify also a scale at which this Ansatz holds, since, even if true at one scale, it will not continue to hold at lower energies. Almost all investigations of supersymmetry phenomenology assume such a degeneracy at a large-energy scale, typically the reduced Planck mass, $M_{\rm p}$. It is often said that degeneracy is automatic in supergravity models, so this is frequently called the supergravity ("SUGRA") model but, as we will see,

supergravity by itself makes *no* prediction of degeneracy. Some authors, similarly, include this assumption as part of the definition of the MSSM, but in this text we will use this term to refer to the particle content and the renormalizable interactions. In any case, the ansatz consists of the following statements at the high-energy scale.

(1) All of the scalar masses are the same, $\widetilde{m}^2 = m_0^2$. This assumption is called *universality* of scalar masses.
(2) The gaugino masses are the same, $M_i = M_0$. This is referred to as the "GUT" relation, since it holds in simple grand unified models.
(3) The soft-breaking cubic terms are assumed to be given by

$$\mathcal{L}_{\text{tri}} = A(H_U Q y_u \bar{u} + H_D Q y_d \bar{d} + H_D L y_l \bar{e}). \tag{11.12}$$

The matrices y_u, y_d, etc., are the same matrices which appear in the Yukawa couplings. This is the assumption of *proportionality*.

Note that with this ansatz, if we ignore possible phases, five parameters are required to specify the model (m_0^2, M_0, A, B_μ, μ). One of these can be traded for M_Z, so this is quite an improvement in predictive power. In addition, this Ansatz automatically satisfies all constraints from rare processes. As we will explain, rare decays and flavor violation are suppressed ($b \to s + \gamma$ is not as strong a constraint, but it requires other relations among soft masses). On the other hand, we will want to ask: just how plausible are these assumptions? We will try to address this question later.

11.1.1 Cancellation of quadratic divergences in gauge theories

We have already seen that soft supersymmetry-violating mass terms receive only logarithmic divergences. While not essential to our present discussion, it is perhaps helpful to see how the cancellation of quadratic divergences for scalar masses arises in gauge theories like the MSSM.

Take, first, a $U(1)$ theory, with (massless) chiral fields ϕ^+ and ϕ^-. Before doing any computation, it is easy to see that provided we work in a way which preserves supersymmetry, there can be no quadratic divergence. In the limit that the mass term vanishes, the theory has a chiral symmetry under which ϕ^+ and ϕ^- rotate by the same phase,

$$\phi^\pm \to e^{i\alpha}\phi^\pm. \tag{11.13}$$

This symmetry forbids a mass term in the superpotential, $\Lambda\phi^+\phi^-$, the only way a supersymmetric mass term could appear. The actual diagrams we need to compute are shown in Fig. 11.1. Since we are interested only in the mass, we can take the

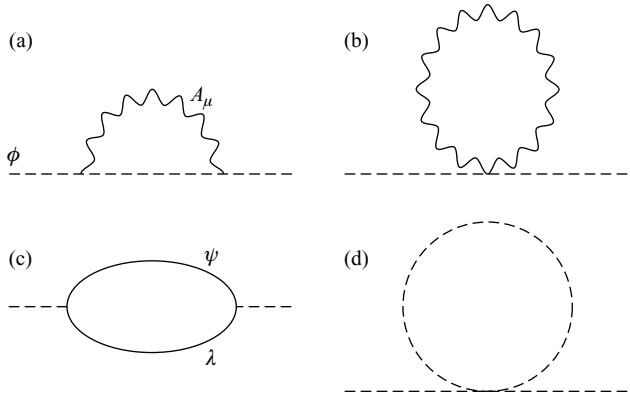

Fig. 11.1. One-loop diagrams contributing to scalar masses in a supersymmetric gauge theory.

external momentum to be zero. It is convenient to choose Landau gauge for the gauge boson. In this gauge the gauge boson propagator is

$$D_{\mu\nu} = -i\left(g_{\mu\nu} - \frac{q_\mu q_\nu}{q^2}\right)\frac{1}{q^2} \tag{11.14}$$

so the first diagram vanishes. The second, third and fourth are straightforward to work out from the basic Lagrangian. One finds:

$$I_b = g^2(i)(i)\frac{3}{(2\pi)^4}\int \frac{d^4k}{k^2}, \tag{11.15}$$

$$I_c = g^2(i)(i)\frac{(\sqrt{2})^2}{(2\pi)^4}\int \frac{d^4k}{k^4}\mathrm{tr}(k_\mu \sigma^\mu k_\nu \bar{\sigma}^\nu) \tag{11.16}$$

$$= -\frac{4g^2}{(2\pi)^4}\int \frac{d^4k}{k^2}, \tag{11.17}$$

$$I_d = g^2(i)(i)\frac{1}{(2\pi)^4}\int \frac{d^4k}{k^2}. \tag{11.18}$$

It is easy to see that the sum, $I_a + I_b + I_c + I_d = 0$.

Including a soft-breaking mass for the scalars, only I_d changes:

$$\begin{aligned} I_d &\to \frac{g^2}{(2\pi)^4}\int \frac{d^4k}{k^2 - \widetilde{m}^2} \\ &= -i\frac{g^2}{(2\pi)^4}\int \frac{d^4k_E}{k_E^2 + \tilde{m}^2} \\ &= \widetilde{m}^2_{\text{independent}} + \frac{ig^2}{16\pi^2}\widetilde{m}^2 \ln(\Lambda^2/\widetilde{m}^2). \end{aligned} \tag{11.19}$$

We have worked here in Minkowski space, and have indicated factors of i to assist the reader in obtaining the correct signs for the diagrams. In the second line, we have performed a Wick rotation. In the third, we have separated off a mass-independent part, since we know that this is canceled by the other diagrams.

Summarizing, the one-loop mass shift is

$$\delta\widetilde{m}^2 = -\frac{g^2}{16\pi^2}\widetilde{m}^2 \ln(\Lambda^2/\widetilde{m}^2). \tag{11.20}$$

Note that the mass shift is proportional to $\widetilde{m}^2$, the supersymmetry breaking mass, as we would expect since supersymmetry is restored as $\widetilde{m}^2 \to 0$. In the context of the Standard Model, we see that the scale of supersymmetry breaking cannot be much larger than the Higgs mass scale itself without fine tuning. Roughly speaking, it can't be much larger than this scale by factors of order $1/\sqrt{\alpha_W}$, i.e., factors of order six. We also see that the correction has a logarithmic sensitivity to the cutoff. So, just like the gauge and Yukawa couplings, the soft masses run with energies.

11.2 $SU(2) \times U(1)$ breaking

In the MSSM, there are a number of general statements which can be made about the breaking of $SU(2) \times U(1)$. The only quartic couplings of the Higgs fields arise from the $SU(2)$ and $U(1)$ D^2 terms. The general form of the soft-breaking mass terms has been described above. So, before worrying about any detailed Ansatz for the soft breakings, the Higgs potential is given, quite generally, by

$$\begin{aligned} V_{\text{Higgs}} = {} & m_{H_U}^2 |H_U|^2 + m_{H_D}^2 |H_D|^2 - m_3^2 (H_U H_D + \text{h.c.}) \\ & + \frac{1}{8}(g^2 + g'^2)(|H_U|^2 - |H_D|^2)^2 + \frac{1}{2} g^2 |H_U H_D|^2. \end{aligned} \tag{11.21}$$

This potential by itself conserves CP; a simple field redefinition removes any phase in m_3^2. (As we will discuss shortly, there are many other possible sources of CP violation in the MSSM.) The physical states in the Higgs sector are usually described by assuming that CP is a good symmetry. In that case, there are two CP-even scalars, H^0 and h^0, where, by convention, h^0 is the lighter of the two. There are a CP-odd neutral scalar, A, and charged scalars, $H^\pm$. At tree level, one also defines a parameter

$$\tan(\beta) = \frac{|\langle H_U \rangle|}{|\langle H_D \rangle|} \equiv v_1/v_2. \tag{11.22}$$

Note that with this definition, as $\tan(\beta)$ grows, so does the Yukawa coupling of the b quark.

To obtain a suitable vacuum, there are two constraints which the soft breakings must satisfy.

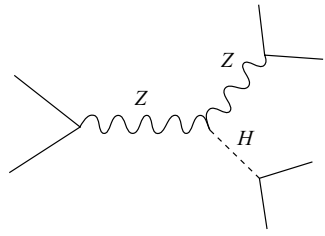

Fig. 11.2. Leading contribution to Higgs production in $e^+ e^-$ annihilation.

(1) Without the soft-breaking terms, $H_U = H_D$ ($v_1 = v_2 = v$) makes the $SU(2)$ and $U(1)$ D terms vanish, i.e. there is no quartic coupling in this direction. So the energy is unbounded below unless

$$m_{H_U}^2 + m_{H_D}^2 - 2|m_3|^2 > 0. \tag{11.23}$$

(2) In order to obtain symmetry breaking, the Higgs mass matrix must have a negative eigenvalue. This gives the requirement:

$$\left|m_3^2\right|^2 > m_{H_U}^2 m_{H_D}^2. \tag{11.24}$$

When these conditions are satisfied, it is straightforward to minimize the potential and determine the spectrum. One finds that

$$m_A^2 = \frac{m_3^2}{\sin(\beta)\cos(\beta)}. \tag{11.25}$$

It is conventional to take m_A^2 as one of the parameters. Then one finds that the charged Higgs masses are given by

$$m_{H^\pm}^2 = m_W^2 + m_A^2, \tag{11.26}$$

while the neutral Higgs masses are:

$$m_{H^0,h^0}^2 = \frac{1}{2}\left(m_A^2 + m_Z^2 \pm \sqrt{\left(m_A^2 + m_Z^2\right)^2 - 4m_Z^2 m_A^2 \cos(2\beta)}\right). \tag{11.27}$$

Note the inequalities:

$$\begin{aligned} m_{h^0} &\leq m_A \\ m_{h^0} &\leq m_Z \\ m_{H^\pm} &\geq m_W. \end{aligned} \tag{11.28}$$

The middle relation is particularly interesting since LEP II set a limit on the Higgs mass of 115 GeV. The basic process through which one searches for the Higgs in e^+e^- annihilation uses the Z–Z–h vertex and exploits a virtual Z, to avoid using the tiny electron Yukawa coupling, as in Fig. 11.2. This limit is actually appropriate

to the Minimal Standard Model, but the limit in the MSSM is not significantly different. Thus it would appear that the MSSM is ruled out. However, these are tree level relations. We will turn shortly to the issue of radiative corrections, and will see that these can be quite substantial – LEP II was not able to rule out the MSSM.

11.3 Why is one Higgs mass negative?

Within the simple ansatz, there is a natural way to understand why $m_{H_U}^2 < 0$ while $m_{H_D}^2 > 0$. What is special about H_U is that it has an $\mathcal{O}(1)$ coupling to the top quark. (If $\tan(\beta)$ is very large, of order 40–50, H_D has a comparable coupling to the b-quark). We saw earlier in the Wess–Zumino model that, at one loop, there is a negative renormalization of the soft-breaking scalar masses. This calculation can be translated to the MSSM, with a modification for the color and $SU(2)$ factors. One obtains:

$$m_{H_U}^2 = (m_{H_U})_0^2 - \frac{6y_t^2}{16\pi^2}\ln(\Lambda^2/m^2)\left(\widetilde{m}_t^2 + m_{\tilde{t}}^2\right), \tag{11.29}$$

$$\widetilde{m}_t^2 = (\widetilde{m}_t)_0^2 - \frac{4y_t^2}{16\pi^2}\ln(\Lambda^2/m^2)\widetilde{m}_H^2. \tag{11.30}$$

So we see that loop corrections involving the top quark Yukawa coupling reduce both the Higgs and the stop masses, but the reduction is larger for the Higgs. If $\Lambda \sim M_{\rm p}$, and the typical soft breakings are of order 1 TeV, these corrections are $\mathcal{O}(1)$, so one needs a full renormalization group analysis to determine if $SU(2) \times U(1)$ is broken. For this we need the full set of renormalization-group equations. These can be derived along the lines of the calculations we have already presented for the gauge and Yukawa contributions to the soft-mass renormalizations. If only y_t is large, the one-loop expressions can be written rather compactly:

$$\mu\frac{\partial}{\partial\mu}m_{H_U}^2 = \frac{1}{8\pi^2}\left[3y_t^2\left(m_{H_U}^2 + m_{\tilde{t}}^2 + m_{Q_3}^2 + \left|A_U^{33}\right|^2\right)\right] - \frac{1}{2\pi^2}\left[|M_2|^2 g_2^2 + |M_1|^2 g_1^2\right] \tag{11.31}$$

$$\mu\frac{\partial}{\partial\mu}m_{\tilde{t}}^2 = \frac{1}{8\pi^2}\left[2y_t^2\left(m_{H_U}^2 + m_{\tilde{t}}^2 + m_{Q_3}^2 + \left|A_U^{33}\right|^2\right)\right] - \frac{1}{2\pi^2}\left[|M_3|^2 g_3^2 + |M_1|^2 g_1^2\right] \tag{11.32}$$

$$\mu\frac{\partial}{\partial\mu}m_{Q_3}^2 = \frac{1}{8\pi^2}\left[2y_t^2\left(m_{H_U}^2 + m_{\tilde{t}}^2 + m_{Q_3}^2 + \left|A_U^{33}\right|^2\right)\right] - \frac{1}{2\pi^2}\left[|M_3|^2 g_3^2 + |M_2|^2 g_2^2 + |M_1|^2 g_1^2\right]. \tag{11.33}$$

Two- (and in some cases higher-) loop corrections to these equations have been computed.

For scalars besides H_U and the third-generation squarks one has only the contribution from diagrams involving intermediate gauginos

$$\mu \frac{\partial}{\partial \mu} m_i^2 = -\frac{1}{2\pi^2} \sum_a g_a^2 |M_a|^2 c_{ai}, \tag{11.34}$$

where c_{ai} denotes the appropriate Casimir (1/2 for particles in the fundamental representation) while gaugino masses satisfy

$$\mu \frac{\partial}{\partial \mu} \left(\frac{M_a^2}{g_a^2} \right) = 0. \tag{11.35}$$

It is straightforward to integrate these equations numerically. For a significant range of parameters, one does obtain suitable breaking of $SU(2) \times U(1)$. But without analyzing these equations in great detail, one can see that, given the large value of m_t, there is a potential fine-tuning problem. Taking Eq. (11.30) as an approximate solution, noting that $y_t \sim 1$ (its precise value depends on $\tan(\beta)$), and allowing for the two types of quark which appear, the coefficient of the logarithm is about $1/8$. At some point, the mass of the Higgs crosses zero. If, at that point, the squark masses are, says 300 GeV^2, then the Higgs mass will become of order 100 GeV almost immediately. So the zero crossing must be almost exactly at the scale of the susy-breaking masses.[1]

11.4 Radiative corrections to the Higgs mass limit

Despite the fact that we know M_H^2 is significantly larger than M_Z^2, this does not (as of this writing) rule out the MSSM. The actual upper limit on the Higgs mass is significantly above its tree-level value, M_Z. At tree level, the form of the Higgs potential is highly constrained. The quartic terms are exactly known. Once supersymmetry is broken, however, there can be corrections to the quartic terms from radiative corrections.[2] These corrections are soft, in that the susy-violating four-point functions vanish rapidly at momenta above the supersymmetry-breaking scale. Still, they are important in determining the low-energy properties of the theory, such as the Higgs vevs and the spectrum.

The largest effect of this kind comes from loops involving top quarks. It is not hard to get a rough estimate of the effect. In the limit $\widetilde{m}_t \gg m_t$, the effective Lagrangian is not supersymmetric below $\tilde{m}_t$. As a result, there can be corrections to the Higgs quartic couplings. Consider the diagrams of Fig. 11.3. In this limit,

[1] This version of the fine-tuning problem was elaborated by N. Arkani-Hamed.

[2] In sufficiently complicated models, there can be tree-level corrections to the quartic couplings. This does not occur in the MSSM, but it can occur in models with singlets.

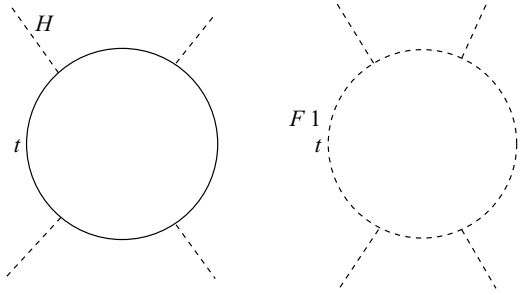

Fig. 11.3. Corrections to quartic Higgs couplings from top loops.

we can get a reasonable estimate by just keeping the top quark loop, cutting off the integral at the mass of the stop. The result will be logarithmically divergent, and we can take the cutoff to be $\widetilde{m}_t$. So we have

$$\delta\lambda = (-1)y_t^4 \times 3 \int \frac{d^4k}{(2\pi)^4} \mathrm{Tr} \frac{1}{(\not{k} - m_t)^4} \tag{11.36}$$

$$= -\frac{12iy_t^4}{16\pi^2} \ln\left(\widetilde{m}_t^2/m_t^2\right). \tag{11.37}$$

The implications of this result are left for the exercises at the end of this chapter.

11.5 Embedding the MSSM in supergravity

In the previous chapter, we introduced $N = 1$ supergravity theories. These theories are not renormalizable, and must be viewed as effective theories, valid below some energy scale, which might be the Planck scale or unification scale (or something else).

The approach we introduced to model building is quite useful as a model for the origin of supersymmetry breaking in the MSSM. The basic assumptions were as follows.

- The theory consists of two sets of fields, the *visible sector fields*, y_a, which in the context of the MSSM would be the quarks and lepton superfields, and the *hidden sector fields*, z_i, responsible for supersymmetry breaking.
- The superpotential was taken to have the form:

$$W(z, y) = W_z(z) + W_y(y). \tag{11.38}$$

- For the Kahler potential we took the simple Ansatz:

$$K = \sum_a y_a^\dagger y_a + \sum_i z_i^\dagger z_i. \tag{11.39}$$

In this case, we saw that if the supersymmetry-breaking scale was of order

$$M_{\rm int} = m_{3/2} M_{\rm p}, \tag{11.40}$$

then there were an array of soft-breaking terms of order $m_{3/2}$. In particular, there were universal masses and A terms,

$$a\, m_{3/2}^2 |y_a|^2 + b\, m_{3/2} W_{ab} y_a y_b + c\, m_{3/2} W_{abc} y_a y_b y_c. \tag{11.41}$$

Here $W_{ab} = \partial_a \partial_b W$; $W_{abc} = \partial_a \partial_b \partial_c W$.

Given that the theory is at best an effective-low-energy theory, one can ask how natural are our assumptions, and what would be the consequences of relaxing them? The assumption that there is some sort of hidden sector, and that the superpotential breaks up as we have hypothesized, we will see, is a reasonable one. It can be enforced by symmetries. The assumption that the Kahler potential takes this simple (often called "minimal") form is a strong one, not justified by symmetry considerations. It turns out not to hold, in any general sense, in string theory, the only context in which presently we can compute it. If we relax this assumption, we lose the universality of scalar masses and the proportionality of the A terms to the superpotential. As we will see in the next section, without these, or something close, the MSSM is not compatible with experiment.

11.6 The μ term

One puzzle in the MSSM is the μ term, the supersymmetric mass term for the Higgs fields. In general, this term is not forbidden by any symmetry, so the first question is: why is it small, of order TeV rather than, say, $M_{\rm p}$ or $M_{\rm gut}$? One possibility is that there is a symmetry. Another is related to the non-renormalization theorems. If, for some reason, there is no mass term at lowest order for the Higgs fields, one will not be generated perturbatively. The μ term, then, might be the result of the same non-perturbative dynamics, for example, responsible for supersymmetry breaking. In string theories, as we will see later, it is quite common to find massless particles at tree level, simply by "accident."

In such a situation, supersymmetry breaking can, quite easily, generate a μ term of order $m_{3/2}$. Consider, for example, the Polonyi model. The operator

$$\int d^4\theta \frac{1}{M_{\rm p}} Z^\dagger H_U H_D \tag{11.42}$$

would generate a μ term of just the correct size. In simple grand unified theories, such a term is often generated.

When we discuss other models for supersymmetry breaking, such as gauge mediation, we will see that the μ term can be more problematic.

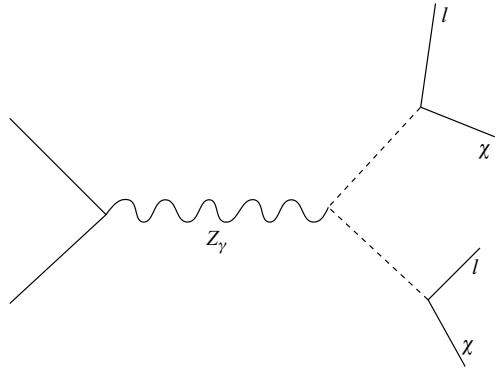

Fig. 11.4. Slepton production in $e^+ e^-$ annihilation.

11.7 Constraints on soft breakings

While the number of soft-breaking parameters is large, they are subject to many experimental constraints. These come from the failure of direct searches to see superpartners of ordinary fields, and also from indirect effects.

11.7.1 Direct searches for supersymmetric particles

While have discussed limits on Higgs particles, there are many other states, and much effort has been devoted at LEP II and the Tevatron to searching for them. The limits are quite impressive. Among the states in the MSSM which are possible discovery channels for supersymmetry, are the "charginos," linear combinations of the partners of the $W^\pm$ and $H^\pm$, and the neutralinos, linear combinations of the partners of the Z and γ (B and W^3) and the neutral Higgs. The mass matrix for the charginos, denoted $w^\pm$ and $\tilde{h}^\pm$, is given by

$$\mathcal{M}_{\chi^\pm} = \begin{pmatrix} M_2 & g v_1 \\ g v_2 & \mu \end{pmatrix}. \tag{11.43}$$

For the neutralinos, w^0, b, $\tilde{h}^0_U$, $\tilde{h}^0_D$, there is a 4×4 mass matrix. We will leave the study of these for the exercises. The lightest of these states is a natural dark matter candidate.

Direct searches at LEP II and the Tevatron strongly constrain the masses of squarks, sleptons, charginos, neutralinos and gluinos. The direct searches are easy to describe, and production and decay rates can be computed given knowledge of the spectrum, since the couplings of the fields are known. If R-parity is conserved, the LSP is stable and weakly interacting, so the characteristic signal for supersymmetry is *missing energy*. For example, in e^+e^- colliders, one can produce slepton pairs, if they are light enough, through the diagram of Fig. 11.4. These then decay, typically,

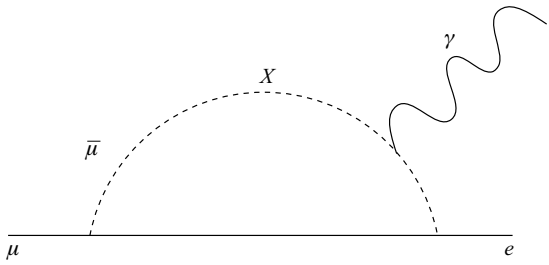

Fig. 11.5. Contribution to $\mu \to e\gamma$.

to a lepton and a neutralino, as indicated. So the final state contains a pair of acoplanar leptons, and missing energy. From such processes, one has (as of this writing) limits of order M_Z or larger, for all of the charged states of the MSSM. One can obtain stronger limits – and in many cases greater discovery potential – from hadron machines. For example, because they are strongly coupled and they are octets of color, gluinos have very substantial production cross sections in hadron collisions. They can be produced both by $q\bar{q}$ and gg annihilation. Gluinos can decay to a large number of channels, and many of these are used in searching for (and setting limits on) the gluino and squark masses. For example, if squarks are heavier than gluinos, gluinos decay through a diagram involving a virtual squark to a pair of quarks and a chargino or neutralino; this particle, in turn, may decay to the LSP and a pair of leptons. This is a quite distinctive signal. The limits on gluinos are greater than 200 GeV.

11.7.2 Constraints from rare processes

Rare processes provide another set of strong constraints on the soft-breaking parameters. In the simple ansatz all of the scalar masses are the same at some very-high-energy scale. However, if this is true at one scale, it is not true at all scales, i.e. these relations are *renormalized*. Indeed, all 105 parameters are truly parameters, and it is not obvious that the assumptions of universality and proportionality are *natural*. On the other hand, there are strong experimental constraints which suggest some degree of degeneracy.

As one example, there is no reason, a priori, why the mass matrix for the $\tilde{L}$s (the partners of the lepton doublets) should be diagonal in the same basis as the charged leptons. If it is not, there is no conservation of separate lepton numbers, and the decay $\mu \to e\gamma$ will occur (Fig. 11.5). To see that we are potentially in serious trouble, we can make a crude estimate. The muon lifetime is proportional

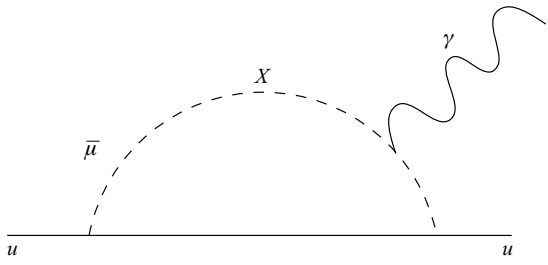

Fig. 11.6. Contribution to $d_{\rm n}$ in supersymmetric theories.

to $G_F^2 m_\mu^5$. The decay, $\mu \to e\gamma$ occurs due to the operator

$$\mathcal{L}_{\mu e\gamma} = eCF_{\mu\nu}\bar{u}\sigma^{\mu\nu}e. \tag{11.44}$$

If there is no particular suppression, we might expect

$$C = \frac{\alpha_w}{\pi}\frac{m_\mu}{m_{\rm susy}^2}. \tag{11.45}$$

So the ratio of these rates would be of order

$$BR = \left(\frac{\alpha_w}{\pi}\right)^2 \left(\frac{M_W}{M_{\rm susy}}\right)^4. \tag{11.46}$$

This might be as small as $10^{-8} - 10^{-9}$ if the supersymmetry-breaking scale is large, 1 TeV or so. But the current experimental limit is 1.2×10^{-11}. So even in this case, it is necessary to suppress the off-diagonal terms. More detailed descriptions of the limits are found in the material in the suggested reading at the end of the chapter.

Another troublesome constraint arises from the neutron and electron electric dipole moments, $d_{\rm n}$ and $d_{\rm e}$. Any non-zero value of these quantities signifies CP-violation. Currently, one has $d_{\rm n} \leq 10^{-25} e$ cm, and $d_{\rm e} \leq 1.6 \times 10^{-27} e$ cm. The soft-breaking terms in the MSSM contain many new sources of CP-violation. Even with the assumptions of universality and proportionality, the gaugino mass and the A, μ and B parameters all are complex, and can violate CP. At the quark level, the issue is that one-loop diagrams can generate a quark dipole moment, as in Fig. 11.6. Note that this particular diagram is proportional to the phases of the gluino and the A parameter. It is easy to see that even if $m_{\rm susy} \sim 500\,{\rm GeV}$, these phases must be smaller than about 10^{-2}. More detailed estimates can be found in the suggested reading at the end of the chapter.

CP is violated in the real world, so it is puzzling that all of the soft-supersymmetry-violating terms should preserve CP to such a high degree. This is in contrast to the Minimal Standard Model, with a single Higgs field, which can reproduce the observed CP-violation with phases of order 1. It is thus a serious

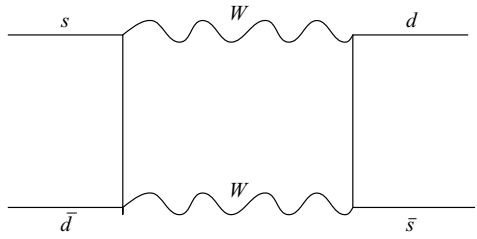

Fig. 11.7. Contribution to $K \leftrightarrow \bar{K}$ in the Standard Model.

challenge to understand why CP should be such a good symmetry if nature is supersymmetric. Various explanations have been offered. We will discuss some of these later, but it should be kept in mind that the smallness of CP violation suggests that either the low-energy supersymmetry hypothesis is wrong, or that there is some interesting physics which explains the surprisingly small values of the dipole moments.

So far, we have discussed constraints on slepton degeneracy and CP-violating phases. There are also constraints on the squark masses arising from various flavor-violating processes. In the Standard Model, the most famous of these are strangeness changing processes, such as K–$\bar{K}$ mixing. One of the early triumphs of the Standard Model was that it successfully explained why this mixing is so small. Indeed, the Standard Model gives a quite good estimate for the mixing. This was originally used to predict – amazingly accurately – the charmed quark mass. The mixing receives contributions from box diagrams such as the one shown in Fig. 11.7. If we consider first, only the first two generations and ignore the quark masses (compared to M_W), we have that

$$\mathcal{M}(K^0 \to \bar{K}^0) \propto (V_{di} V_{is}^\dagger)(V_{sj}^\dagger V_{jd}) = 0. \tag{11.47}$$

Including fermion masses leads to terms in $\mathcal{L}_{\text{eff}}$ of order

$$\frac{\alpha_W}{4\pi} \frac{m_c^2}{M_W^2} G_{\text{F}} \ln\left(m_c^2/m_u^2\right)(\bar{s}\gamma^\mu\gamma_5 d)(\bar{d}\gamma^\mu\gamma^5 s) + \cdots. \tag{11.48}$$

The matrix element of the operator appearing here can be estimated in various ways, and one finds that this expression roughly saturates the observed value (this was the origin of the prediction by Gaillard and Lee of the value of the charmed quark mass). Similarly, the CP-violating part (the "ϵ" parameter) is in rough accord with observation, for reasonable values of the KM parameter δ.

In supersymmetric theories, if squarks are degenerate, there are similar cancellations. However, if they are not, there are new, very dangerous contributions. The most serious is that indicated in Fig. 11.8, arising from exchange of gluinos

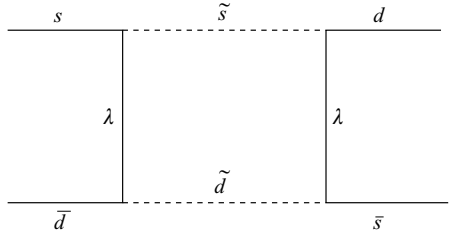

Fig. 11.8. Gluino exchange contribution to kaon mixing in the MSSM.

and squarks. This is nominally larger than the Standard Model contribution by a factor of $(\alpha_s/\alpha_W)^2 \approx 10$. Also, the Standard Model contribution vanishes in the chiral limit, whereas the gluino exchange does not, and this leads to an additional enhancement of nearly an order of magnitude. On the other hand, the diagram is highly suppressed in the limit of exact universality and proportionality. Proportionality means that the A terms in Eq. (11.8) are suppressed by factors of light quark masses, while universality means that the squark propagator, $\langle \tilde{q}^* \tilde{q} \rangle$, is proportional to the unit matrix in flavor space. So there are no appreciable off-diagonal terms which can contribute to the diagram. On the other hand, there is surely some degree of non-degeneracy. One finds that even if the characteristic susy scale is 500 GeV, one needs degeneracy in the down-squark sector at the part in 10^2 level.

So K–$\overline{K}$ mixing tightly constrains the down-squark mass matrix. The imaginary part provides further constraints. There are also strong limits on D–$\overline{D}$ mixing, which significantly restrict the mass matrix in the up-squark sector. Other important constraints on soft breakings come from other rare processes, such as $b \to s\gamma$. Again, more details can be found in the references in the suggested reading.

Suggested reading

The MSSM is described in most reviews of supersymmetry. Probably the best place to look for up-to-date reviews of the model and the experimental constraints is the Particle Data Group website. A useful collection of renormalization group formulas for supersymmetric theories is provided in the review by Martin and Vaughn (1994). Limits on rare processes are discussed in a number of articles, such as that by Masiero and Silvestrini (1997).

Exercises

(1) Verify the expressions for $I_a - I_d$ (Eqs. (11.15)–(11.18)).
(2) Derive Eqs. (11.23)–(11.26).

(3) Verify the formula for the top quark corrections to the Higgs mass. Evaluate y_t in terms of m_t and $\sin(\beta)$. Show that, to this level of accuracy,

$$m_h^2 < m_Z^2 \cos(2\beta) + \frac{12g^2}{16\pi^2}\frac{m_t^4}{m_W^2}\ln\left(\widetilde{m}^2 m_t^2\right).$$

(4) Estimate the size of the supersymmetric contributions to the quark electric dipole moment, assuming that all of the superpartner masses are of order m_{susy}, and δ is a typical phase. Assuming, as well, that the neutron electric dipole moment is of order the quark electric dipole moment, how small do the phases have to be if $m_{\text{susy}} = 500\,\text{GeV}$?

12

Supersymmetric grand unification

In this brief chapter, we discuss one of the most compelling pieces of circumstantial evidence in favor of supersymmetry: unification of coupling constants. Earlier, we introduced grand unification without supersymmetry. In this chapter, we consider how supersymmetry modifies that story.

12.1 A supersymmetric grand unified model

Just as in theories without supersymmetry, the simplest group into which one can unify the gauge group of the Standard Model is $SU(5)$. The quark and lepton superfields of a single generation again fit naturally into a $\bar{5}$ and 10.

To break $SU(5)$ to $SU(3) \times SU(2) \times U(1)$, we can again consider a 24 of Higgs fields, Σ. If we wish supersymmetry to be unbroken at high energies, the superpotential for this field should not lead to supersymmetry breaking. The simplest renormalizable superpotential is:

$$W(\Sigma) = m\mathrm{Tr}\Sigma^2 + \frac{\lambda}{3}\mathrm{Tr}\Sigma^3. \tag{12.1}$$

Treating this as a globally supersymmetric theory (i.e. ignoring supergravity corrections), the equations

$$\frac{\partial W}{\partial \Sigma} = 0 \tag{12.2}$$

are conveniently studied by introducing a Lagrange multiplier to enforce $\mathrm{Tr}\Sigma = 0$. The resulting equations have three solutions:

$$\Sigma = 0; \Sigma = \frac{m}{\lambda} \quad \mathrm{diag}(1, 1, 1, -4); \quad \Sigma = \frac{m}{\lambda}\mathrm{diag}(2, 2, 2, -3, -3). \tag{12.3}$$

These solutions either leave $SU(5)$ unbroken, or break $SU(5)$ to $SU(4) \times U(1)$ or to the Standard Model group. Each of these solutions is isolated; you can check

that there are no massless fields from Σ in any of these states. At the classical level, they are degenerate.

If we include supergravity corrections, these states are split in energy. Provided the unification scale, m, is substantially below the Planck scale, these corrections can be treated perturbatively. In order to make the cosmological constant vanish in the $SU(3) \times SU(2) \times U(1)$ ((3, 2, 1)) vacuum, it is necessary to include a constant in the superpotential, such that, in this vacuum, the expectation value of the superpotential is zero. As a result, the other two states have negative energy (as we will see in the chapter on gravitation, they correspond to solutions in which space-time is not Minkowski, but *anti-de Sitter*).

We will leave working out the details of these computations to the exercises, and turn to other features of this model. It is necessary to include Higgs fields to break $SU(2) \times U(1)$ down to $U(1)$. The simplest choice for the Higgs representation is the 5. As in the MSSM, it is actually necessary to introduce two sets of fields so as to avoid anomalies: a 5 and $\bar{5}$ are the minimal choice. We denote these fields by H and $\bar{H}$.

Once again, it is important that the color triplet Higgs fields in these multiplets be massive in the (3, 2, 1) vacuum. The most general renormalizable superpotential coupling the Higgs to the adjoint is:

$$m_H H \bar{H} + y \bar{H} \Sigma H. \tag{12.4}$$

By carefully adjusting y (or m), we can arrange that the Higgs doublet is massless. As a result, the triplet is automatically massive, with mass of order m_H. Of course, this represents an extreme fine tuning. We will see that the unification scale is about 10^{16} GeV, so this is a tuning of a part in 10^{13} or so. But it is curious that this tuning only need be done classically. Because the superpotential is not renormalized, radiative corrections do not lead to large masses for the doublets.

12.2 Coupling constant unification

The calculation of coupling constant unification in supersymmetric theories is quite similar to that in non-supersymmetric ones. We assume that the threshold for the supersymmetric particles is somewhere around 1 TeV. So up to that scale, we run the renormalization group equations just as in the Standard Model. Above that scale, there are new contributions from the superpartners of ordinary particles. The leading terms in the beta functions are:

$$\mathrm{SU(3)} : b_0 = 3; \quad \mathrm{SU(2)} : b_0 = -1; \quad \mathrm{U(1)} : b_0 = -33/5. \tag{12.5}$$

One can be more thorough, including two loop corrections and threshold effects. The result of such an analysis are shown in Fig. 12.1. One has:

$$M_{\mathrm{gut}} = 1.2 \times 10^{16} \mathrm{GeV}; \quad \alpha_{\mathrm{gut}} \approx \frac{1}{25}. \tag{12.6}$$

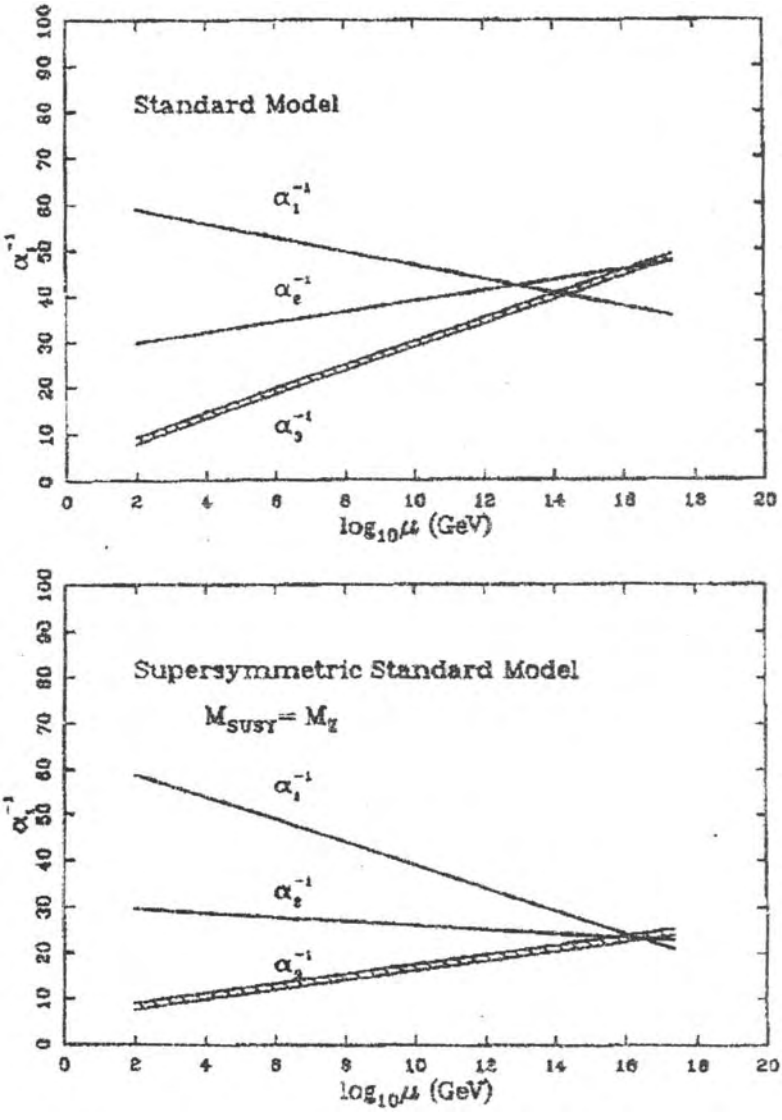

Fig. 12.1. In the Standard Model, the couplings do not unify at a point. In the MSSM, they do, provided that the threshold for new particle production is at about 1 TeV. Reprinted with permission from P. Langacker and N. Polonsky. Uncertainties in coupling constant unification. *Phys. Rev. D*, **47**, 4028 (1993). Copyright (1993) by the American Physical Society.

The agreement in the figure is striking. One can view this as a successful prediction of α_s, given the values of the $SU(2)$ and $U(1)$ couplings.

12.3 Dimension-five operators and proton decay

We have seen that in supersymmetric theories, there are dangerous dimension four operators. These can be forbidden by a simple Z_2 symmetry, R-parity. But there are also operators of dimension five which can potentially lead to proton decay rates far larger than the experimental limits. The MSSM possesses B- and L-violating dimension-five operators which are permitted by all symmetries. For example, R-parity doesn't forbid such operators as

$$\mathcal{O}_5^a = \frac{1}{M}\int d^2\theta \bar{u}\bar{u}\bar{d}e^+ \quad \mathcal{O}_5^b = \frac{1}{M}\int d^2\theta QQQL. \tag{12.7}$$

These are still potentially very dangerous. When one integrates out the squarks and gauginos, they will lead to dimension-six B- and L-violating operators in the Standard Model with coefficients (optimistically) of order

$$\frac{\alpha}{4\pi}\frac{1}{Mm_{\text{susy}}}. \tag{12.8}$$

Comparing with the usual minimal $SU(5)$ prediction, and supposing that $M \sim 10^{16}$ GeV, one sees that one needs a suppression of order 10^9 or so.

Fortunately, such a suppression is quite plausible, at least in the framework of supersymmetric GUTs. In a simple $SU(5)$ model, for example, the operators of Eq. (12.7) will be generated by exchange of the color triplet partners of ordinary Higgs fields, and thus one gets two factors of Yukawa couplings. Also, in order that the operators be $SU(3)$-invariant, the color indices must be completely anti-symmetrized, so more than one generation must be involved. This suggests that suppression by factors of order Cabibbo angles is plausible. So we can readily imagine a suppression by factors of 10^{-9}–10^{-11}. Proton decay can be used to restrict – and does severely restrict – the parameter space of particular models. But what is quite striking is that we are automatically in the right range to be compatible with experimental constraints, and perhaps even to see something. It is not obvious that things had to be this way.

So far we have phrased this discussion in terms of baryon-violating physics at M_{gut}. But whatever the underlying theory at M_{p} may be, there is no reason to think that it should preserve baryon number. So one expects that already at scales just below M_{p}, these dimension-five terms are present. If their coefficients were simply of order $1/M_{\text{p}}$, the proton decay rate would be enormous, five orders of magnitude or more faster than the current bounds. But in any such theory, one

must also explain the smallness of the Yukawa couplings. One popular approach is to postulate approximate symmetries. Such symmetries could well suppress the dangerous operators at the Planck scale. One might expect that there would be further suppression, in any successful underlying theory. After all, the rate from Higgs exchange in GUTs is so small because the Yukawa couplings are small. We do not really know why Yukawa couplings are small, but it is natural to suspect that this is a consequence of (approximate) symmetries. These same symmetries, if present would also suppress dimension-five operators from Planck scale sources, presumably by a comparable amount.

Finally, we mentioned earlier that one can contemplate symmetries to suppress dimension-four operators beyond a Z_2 R-parity. Such symmetries, as we will see, are common in string theory. One can write down R-symmetries which forbid, not only all of the dangerous dimension four operators, but some or all of the dimension-five operators as well. In this case, proton decay could be unobservable in feasible experiments.

Suggested reading

A good introduction to supersymmetric GUTs is provided in Witten (1981). The reviews and texts which we have mentioned on supersymmetry and grand unification all provide good coverage of the topic. The Particle Data Group website has an excellent survey, including up-to-date unification calculations and constraints on dimension-five operators (see Eidelman *et al*. 2004).

Exercises

(1) Work through the details of the simplest $SU(5)$, supersymmetric grand unified model. Solve the equations

$$\frac{\partial W}{\partial \Sigma} = 0.$$

Couple the system to supergravity, and determine the value of the constant in the superpotential required to cancel the cosmological constant in the $(3, 2, 1)$ minimum. Determine the resulting value of the vacuum energy in the $SU(5)$ symmetric minimum.

(2) In the simplest $SU(5)$ model, include a 5 and $\bar{5}$ of Higgs fields. Write the most general renormalizable superpotential for these fields and the 24, Σ. Find the condition on the parameters of the superpotential so that there is a single light doublet. Using the fact that only the Kahler potential is renormalized, show that this tuning of parameters at tree level assures that the doublet remains massless to all orders of perturbation theory.

Now consider the couplings of quarks and leptons required to generate masses for the fermions. Show that exchanges of 5 and $\bar{5}$ Higgs lead to baryon and lepton number violating dimension-five couplings.

(3) Show how various B-violating four-Fermi operators are generated by squark and slepton exchange, starting with the general set of B- and L-violating terms in the superpotential.

13

Supersymmetric dynamics

In the previous chapter, we understood how to build realistic particle-physics models based on supersymmetry. There are already significant constraints on such theories, and experiments at the LHC will test whether these sorts of ideas are correct.

If supersymmetry is discovered, the question will become: how is supersymmetry broken? Supersymmetry breaking offers particular promise for explaining large hierarchies. Consider the non-renormalization theorems. Suppose we have a model consisting of chiral fields and gauge interactions. If the superpotential is such that supersymmetry is unbroken at tree level, the non-renormalization theorems for the superpotential which we proved in Section 9.7 guarantee that supersymmetry is not broken to all orders of perturbation theory. But they do not necessarily guarantee that effects *smaller* than any power of the couplings don't break supersymmetry. So, if we denote the generic coupling constants by g^2, there might be effects of order, say, e^{-c/g^2} which break the symmetry. In the context of a theory like the MSSM, supposing that soft breakings are of this order, this might account for the wide disparity between the weak scale (correlated with the susy-breaking scale) and the Planck or unification scale.

So one reason that the dynamics of supersymmetric theories is of interest is to understand dynamical supersymmetry breaking, and perhaps to study a whole new class of phenomena in nature. But there are other reasons to be interested, as first most clearly appreciated by Seiberg. Supersymmetric Lagrangians are far more tightly constrained than ordinary Lagrangians. It is often possible to make strong statements about dynamics which would be difficult if not impossible to study in conventional field theories. We will see this includes phenomena like electric–magnetic duality and confinement.

13.1 Criteria for supersymmetry breaking: the Witten index

We will consider a variety of theories, many of them strongly coupled. One might imagine that it is a hard problem to decide if supersymmetry is or is not broken. Even in weakly coupled theories, one might wonder whether one could establish reliably that supersymmetry is *not* broken, since, unless one solved the theory exactly, it would seem hard to assert that there was no tiny non-perturbative effect which did not break the symmetry. One of the things we will learn in this chapter is that this is not a particularly difficult problem. We will exploit several tools. One is known as the Witten index. Consider the field theory of interest in a finite box. At finite volume, the supersymmetry charges are well defined, whether or not supersymmetry is spontaneously broken. Because of the supersymmetry algebra,

$$Q|B\rangle = \sqrt{E}|F\rangle,\ Q|F\rangle = \sqrt{E}|B\rangle, \tag{13.1}$$

i.e. non-zero-energy states come in Fermi–Bose pairs. Zero-energy states are special; they need not be paired. In the infinite-volume limit, the question of supersymmetry breaking is the question of whether there are or are not zero-energy states. To count these, Witten suggested evaluating:

$$\Delta = \mathrm{Tr}(-1)^F e^{-\beta H}. \tag{13.2}$$

Non-zero-energy states do not contribute to the index. The exponential is present to provide an ultraviolet regulator: Δ is *independent of* β. More strikingly, the index is independent of all of the parameters of the theory. The only way Δ can change as some parameter is changed is by some zero energy state acquiring non-zero energy, or a non-zero-energy state acquiring zero energy. But, because of Eq. (13.1), whenever the number of zero-energy bosonic states changes, the number of zero-energy fermionic states changes by the same amount. The index is thus topological in character, and it is from this that it derives its power, as well as its applications in a number of areas of mathematics. What can we learn from the index? If $\Delta \neq 0$, we can say with confidence that supersymmetry is not broken. If $\Delta = 0$, we don't know.

Let's consider an example: a supersymmetric gauge theory, with gauge group $SU(2)$, and no chiral fields. Since Δ is independent of parameters, we can consider the theory in a very tiny box, with very small coupling. We can evaluate Δ, somewhat heuristically, as follows. Work in $A_0 = 0$ gauge. Consider, first, the bosonic degrees of freedom, the A_is, where the A_is are matrix valued. In order for the energy to be small, we need the A_is to be constant, and to commute. So take A_i to lie in the third dimension in the isospin space, and ignore the other bosonic degrees of freedom. One might try to remove these remaining variables by a gauge transformation, $g = \exp(i A_i x^i)$, but g is only a sensible gauge transformation if it is single-valued,

which means that $A_i^3 = 2\pi n/L$. So A_i^3 is a compact variable. This reduces the problem to the quantum mechanics of a rotor. So in the lowest state the wave function is a constant. Because the A_i^3s are non-zero, the lowest energy states will only involve the gluinos in the 3 direction. There are two of them, λ_1^3 and λ_2^3 (again independent of coordinates).

Now recall that, in $A_0 = 0$ gauge, the states must be gauge invariant. One interesting gauge transformation is multiplication by σ_2. This flips the sign of A^3 and λ^3. If we assume that our Fock ground state is even under this transformation, the only invariant states are $|0\rangle$ and $\lambda_1^3\lambda_2^3|0\rangle$. So we find $\Delta = 2$. If we assume that the state is odd, then we obtain $\Delta = -2$.

As we indicated, this argument is heuristic. A more detailed, but still heuristic, argument was provided by Witten in his paper on the index. But Witten also provided a more rigorous argument, which yields the same result. For general $SU(N)$, one finds $\Delta = N$.

This already establishes that a vast array of interesting supersymmetric field theories do not break supersymmetry: not only all of the pure gauge theories, but any theory with massive matter fields. This follows from the independence of Δ of parameters. If the mass is finite, one can take it to be large; if it is sufficiently large, we can ignore the matter fields and recover the pure gauge result. Later, we will understand the dynamics of these theories in some detail, and will reproduce the result for the index. But we will also see that the limit of zero mass is subtle, and the index calculation is not directly relevant.

13.2 Gaugino condensation in pure gauge theories

Our goal in this section is to understand the dynamics of a pure $SU(N)$ gauge theory, with massless fermions in the adjoint representation. Without thinking about supersymmetry, one might expect the following, from our experience with real QCD.

(1) The theory has a mass gap, i.e. the lowest excitations of the theory are massive.
(2) Gauginos, like quarks, condense, i.e.

$$\langle\lambda\lambda\rangle = c\Lambda^3 = ce^{-(8\pi^2/b_0 g^2)}. \tag{13.3}$$

Note that there is no Goldstone boson associated with the gluino (gaugino) condensate. The theory has no continuous global symmetry; the classical symmetry,

$$\lambda \to e^{i\alpha}\lambda, \tag{13.4}$$

is anomalous. However, a discrete subgroup,

$$\lambda \to e^{\frac{2\pi i}{N}}\lambda, \tag{13.5}$$

is free of anomalies. One can see this by considering instantons in this theory. The instanton has $2N$ zero modes; this would appear to preserve a Z_{2N} symmetry. But the transformation $\lambda \to -\lambda$ is actually equivalent to a Lorentz transformation (a rotation by 2π). Multi-instanton solutions also preserve this symmetry, and it is believed to be exact. So the gaugino condensate breaks the Z_N symmetry; there are N degenerate vacua. This neatly accounts for the N of the index. Later we will show that, even though the theory is strongly coupled, we can demonstrate by a controlled semiclassical computation the existence of the condensate.

Gluino condensation implies a breakdown of the non-renormalization theorems at the non-perturbative level. Recall that the Lagrangian is:

$$\mathcal{L} = \int d^2\theta S W_\alpha^2 \tag{13.6}$$

so $\langle\lambda\lambda\rangle$ gives rise to a superpotential, i.e.

$$\mathcal{L} = \int d^2\theta S\langle\lambda\lambda\rangle. \tag{13.7}$$

This is our first example of a non-perturbative correction to the superpotential. Note, however, that $\langle\lambda\lambda\rangle$ depends on S, since it depends on g^2:

$$S\langle\lambda\lambda\rangle = e^{-\frac{3S}{b_0}}. \tag{13.8}$$

So we actually have a superpotential for S:

$$W(S) = e^{-\frac{S}{N}}. \tag{13.9}$$

This superpotential violates the continuous shift symmetry which we used to prove the non-renormalization theorem, but it is compatible with the non-anomalous R-symmetry:

$$S \to S + i\alpha N, \quad \lambda \to \lambda e^{i\alpha}. \tag{13.10}$$

Under this symmetry, the superpotential transforms with charge 2.

13.3 Supersymmetric QCD

A rich set of theories for study are collectively referred to as "supersymmetric QCD." These are gauge theories with gauge group $SU(N)$, N_f fields, Q_f, in the N representation, and N_f fields, $\bar{Q}_f$, in the $\bar{N}$ representation. We will see that the dynamics is quite sensitive to the value of N_f. First consider the theory without any classical superpotential for the quarks. In this case, the theory has a large global symmetry. We can transform the Qs and $\bar{Q}$s by separate $SU(N)$ transformations. We can also multiply the Qs by a common phase, and the $\bar{Q}$s by a separate common

phase:

$$Q_f \to e^{i\alpha} Q_f \quad \bar{Q}_f \to e^{i\beta} \bar{Q}_f. \tag{13.11}$$

Finally, the theory possesses an R-symmetry, under which the Qs and $\bar{Q}$s are neutral. In terms of component fields, under this symmetry:

$$\psi_Q \to e^{-i\alpha} \psi_Q \quad \psi_{\bar{Q}} \to e^{-i\alpha} \psi_{\bar{Q}} \quad \lambda^a \to e^{i\alpha} \lambda^a. \tag{13.12}$$

Now consider the question of anomalies. The $SU(N)$ symmetries are free of anomalies, as is the vector-like symmetry,

$$Q_f \to e^{i\alpha} Q_f \quad \bar{Q}_f \to e^{-i\alpha} \bar{Q}_f. \tag{13.13}$$

The R-symmetry, and the axial $U(1)$, are both anomalous. But we can define a non-anomalous R by combining the two. The gauginos give a contribution to the anomaly proportional to N; so we need the fermions to carry R-charge $-N/N_{\rm f}$. Since the bosons (and the chiral multiplets) carry R-charge larger by 1, we have

$$Q_f(x,\theta) \to e^{i\alpha \frac{N_{\rm f}-N}{N_{\rm f}}} Q_f(x, \theta e^{-i\alpha}) \quad \bar{Q}_f(x,\theta) \to e^{i\alpha \frac{N_{\rm f}-N}{N_{\rm f}}} \bar{Q}_f(x, \theta e^{-i\alpha}), \tag{13.14}$$

so the symmetry of the quantum theory is $SU(N_{\rm f})_{\rm L} \times SU(N_{\rm f})_{\rm R} \times U(1)_{\rm R} \times U(1)_{\rm V}$.

We have seen that supersymmetric theories often have, classically, a large vacuum degeneracy, and this is true of this theory. In the absence of a superpotential, the potential is completely determined by the D terms for the gauge fields. It is helpful to treat D as a matrix-valued field,

$$D = \sum T^a D^a. \tag{13.15}$$

As a matrix, D can be expressed elegantly in terms of the scalar fields. We start with the identity:

$$(T^a)_i^j (T^a)_k^l = \delta_i^l \delta_k^j - \frac{1}{N} \delta_i^j \delta_k^l. \tag{13.16}$$

One can derive this result a number of ways. Consider propagators for fields (like gauge bosons) in the adjoint representation of the gauge group. Take the group, first, to be $U(N)$. The propagator of the matrix-valued fields

$$\left\langle A_i^j A_k^l \right\rangle \propto \delta_i^l \delta_k^j. \tag{13.17}$$

But this is the same thing as

$$\left\langle A^a A^b (T^a)_i^j (T^b)_k^l \right\rangle. \tag{13.18}$$

So we obtain the identity without the $1/N$ terms. Now remembering that A must be traceless, we see that we need to subtract the trace as above. (This identity

is important in understanding the $1/N$ expansion in QCD.) So a field, ϕ, in the fundamental representation, makes a contribution:

$$\delta D_i^j = \phi_i^* \phi^j - \frac{1}{N} \delta_i^j \phi_k^* \phi^k. \tag{13.19}$$

In the anti-fundamental representation, the generators are $-T^{aT}$ (this follows from the fact that the generators in the anti-fundamental are minus the complex conjugates of those in the fundamental, and the fact that the T^as are Hermitian). So the full D term is:

$$D_i^j = \sum_f Q_i^* Q^j - \bar{Q}_i \bar{Q}^{*\,j} - \text{Tr terms}. \tag{13.20}$$

In this matrix form it is not difficult to look for supersymmetric solutions, i.e. solutions of $D_i^j = 0$. A simple strategy is first to construct

$$\hat{D}_i^j = \sum_f Q_i^* Q^j - \bar{Q}_i \bar{Q}^{*\,j} \tag{13.21}$$

and demand that $\hat{D}$ either vanish or be proportional to the identity. Let's start with the case $N_\mathrm{f} \leq N$. For definiteness, take $N = 3$, $N_\mathrm{f} = 2$; the general case is easy to work out. By a sequence of $SU(3)$ transformations, we can bring Q to the form:

$$Q = \begin{pmatrix} v_{11} & v_{12} \\ 0 & v_{22} \\ 0 & 0 \end{pmatrix}. \tag{13.22}$$

By a sequence of $SU(N_\mathrm{f})$ transformations, we can bring this to a similar form:

$$Q = \begin{pmatrix} v_1 & 0 \\ 0 & v_2 \\ 0 & 0 \end{pmatrix}. \tag{13.23}$$

At this point we have used up our freedom to make further symmetry transformations on $\bar{Q}$. But it is easy to find the most general $\bar{Q}$ which makes the D terms vanish. The contribution of Q to D_i^j is simply

$$D = \mathrm{diag}(|v_1|^2, |v_2|^2). \tag{13.24}$$

So, in order that D vanish, $\bar{Q}$ must make an equal and opposite contribution. In order that there be no off-diagonal contributions, $\bar{Q}$ can have entries only on the diagonal, so

$$\bar{Q} = \begin{pmatrix} e^{i\alpha_1} v_1 & 0 \\ 0 & e^{i\alpha_2} v_2 \\ 0 & 0 \end{pmatrix}. \tag{13.25}$$

In general, in these directions, the gauge group is broken to $SU(N - N_f)$. The unbroken flavor group depends on the values of the v_is. We have exhibited N_f complex moduli above, but actually there are more; these are associated with the generators of the broken flavor symmetries ($SU(N_f) \times U(1)$) so there are $N_f^2 + 2N_f$ complex moduli. Note that there are $2NN_f - N_f^2$ broken gauge generators, which gain mass by "eating" the components of Q, $\bar{Q}$ that are not moduli. This leaves, of the original $2NN_f$ chiral fields, precisely $N_f^2 + 2N_f$ massless fields, so we have correctly identified the number of moduli.

Our discussion, so far, does not look gauge invariant. But this is easily, and elegantly, rectified. The moduli can be written as the gauge-invariant combinations:

$$M_{\bar{f}}^{f} = \bar{Q}_{\bar{f}} Q^{f}. \tag{13.26}$$

Expanding the fields Q and $\bar{Q}$ about their expectation values gives back the explicit form for the moduli in terms of the underlying, gauge-variant fields. This feature, we will see, is quite general.

The case $N_f = N$ is similar to the case $N_f < N$, but there is a significant new feature. In addition to the flat directions with $Q = \bar{Q}$ (up to phases), the potential also vanishes if $Q = v\mathrm{I}$, where I is the identity matrix. This possibility can also be described in a gauge-invariant way, since now we have an additional pair of gauge invariant fields, which we will refer to as "baryons":

$$B = \epsilon^{i_1 \ldots i_N} \epsilon_{a_1 \ldots a_N} Q_{i_1}^{a_1} \ldots Q_{i_N}^{a_N}, \tag{13.27}$$

and similarly for $\bar{B}$.

In the case $N_f > N$, there is a larger set of baryon-like objects, corresponding to additional flat directions. We will describe them in greater detail later. Before closing this section, we should stress that, for $N_f \geq N - 1$, the gauge symmetry is completely broken. For large values of the moduli, the effective coupling of the theory is $g^2(v)$, since infrared physics cuts off at the scale of the gauge field masses. By taking v as large enough that $g^2(v)$ is small, the theories can be analyzed by perturbative and semiclassical methods. Strong coupling is more challenging, but much can be understood. We will see that the dynamics naturally divides into three cases: $N_f < N$, $N_f = N$, and $N_f > N$.

13.4 $N_f < N$: a non-perturbative superpotential

Our problem now is to understand the dynamics of these theories. Away from the origin of the moduli spaces, this turns out to be a tractable problem. We consider first the case $N_f < N$. Suppose that the vs are large and roughly uniform in magnitude. Even here, we have to distinguish two cases. If $N_f = N - 1$, the gauge group is

completely broken; the low-energy dynamics consists of the set of chiral fields, $M_{\bar{f},f}$. If $N_{\rm f} < N-1$, there is an unbroken gauge group, $SU(N-N_{\rm f})$, with no matter fields (chiral fields) transforming under this group at low energies. The gauge theory is an asymptotically free theory, essentially like ordinary QCD with fermions in the adjoint representation. Such a theory is believed to have a mass gap, of order the scale of the theory, $\Lambda_{N-N_{\rm f}}$. Below this scale, again, the only light fields are the moduli $M^{f}_{\bar{f}}$. In both cases, we can try to guess the form of the very-low-energy effective action for these fields from symmetry considerations.

We are particularly interested in whether there is a superpotential in this effective action. If not, the moduli have, *exactly*, no potential. In other words, even in the full quantum theory, they correspond to an exact, continuous set of ground states. What features should this superpotential possess? Most important, it should respect the flavor symmetries of the original theory (because the fields M are gauge invariant, it automatically respects the gauge symmetry). Among these symmetries are the $SU(N_{\rm f}) \times SU(N_{\rm f})$ non-Abelian symmetry. The only invariant we can construct from M is

$$\Phi = \det M. \tag{13.28}$$

The determinant is invariant because it transforms under $M \to VMU$ as $\det V \det U \det M$, and, for $SU(N_{\rm f})$ transformations, the determinant is unity. Under the baryon number symmetry, M is invariant. But, under the $U(1)_R$ symmetry, its transformation law is more complicated:

$$\Phi \to e^{2i\alpha(N_{\rm f}-N)}. \tag{13.29}$$

Under this R-symmetry, any would-be superpotential must transform with charge 2, so the form of the superpotential is unique:

$$W = \Lambda^{(3N-N_{\rm f})/(N-N_{\rm f})}\Phi^{-1/(N-N_{\rm f})}. \tag{13.30}$$

Here we have inserted a factor of Λ, the scale of the theory, on dimensional grounds.

Our goal in the next two sections will be to understand the dynamical origin of this superpotential, known as the ADS (after Affleck, Dine and Seiberg) superpotential. We will see that there is a distinct difference between the cases $N_{\rm f} = N-1$ and $N_{\rm f} < N-1$. First, though, consider the case $N = N_{\rm f}$. Then the field, Φ, has R-charge zero, and no superpotential is possible. So no potential can be generated, perturbatively or non-perturbatively. Similarly, in the case $N_{\rm f} > N$, we cannot construct a gauge-invariant field which is also invariant under the $SU(N_f) \times SU(N_f)$ flavor symmetry. This may not be obvious, since it would seem that we could again construct $\Phi = \det(M)$. But in this case, $\Phi = 0$, by antisymmetry.

From the perspective of ordinary, non-supersymmetric field theories, what we have established here is quite surprising. Normally, we would expect that, in an

interacting theory, even if the potential vanished classically, there would be quantum corrections. For theories with $N \geq N_f$, we have just argued that this is impossible. So this is a new feature of supersymmetric theories: there are often exact moduli spaces, even at the quantum level.

In the next few sections, we will demonstrate that non-perturbative effects do indeed generate the superpotential of Eq. (13.30). The presence of the superpotential means that, at least at weak coupling (large v), there is no stable vacuum of the theory. At best, we can think about time-dependent, possibly cosmological, solutions. If we add a mass term for the quarks, however, we find an interesting result. If the masses are the same, we expect that all of the v_is will be the same, $v_i = v$. Suppose that the mass term is small. Then the full superpotential, at low energies, is

$$W = m\bar{Q}Q + \Lambda^{(3N-N_f)/(N-N_f)}\Phi^{-1/(N-N_f)}. \tag{13.31}$$

Remembering that $\Phi \sim v^{2N_f}$, the equation for a supersymmetric minimum has the form:

$$v^{2N/(N-N_F)} = \left(\frac{m}{\Lambda}\right)\Lambda^{2N/(N-N_F)}. \tag{13.32}$$

Note that v is a complex number; this equation has N roots:

$$v = e^{\frac{2\pi ik}{N}}\Lambda\left(\frac{m}{\Lambda}\right)^{(N-N_f)/2N}. \tag{13.33}$$

What is the significance of these N solutions? The mass term breaks the $SU(N_f) \times SU(N_f)$ symmetry to the vector sum. It also breaks the $U(1)_R$. But it leaves unbroken a Z_N subgroup of the $U(1)$. In Eq. (13.14), $\alpha = 2N_f/N$ is a symmetry of the mass term. So these N vacua are precisely those expected from breaking the Z_N. This Z_N is the same Z_N expected for a pure gauge theory, as one can see by thinking of the case where the mass of the Qs and $\bar{Q}$s is large.

13.4.1 The Λ-dependence of the superpotential

Previously, we proved a non-renormalization theorem for the gauge couplings by thinking of the gauge coupling itself as a background field, S. This relied on the shift symmetry,

$$S \to S + i\alpha.$$

This symmetry, however, is only a symmetry of perturbation theory. Since the imaginary part of S, a, couples to $F\tilde{F}$, instanton and other non-perturbative effects violate the symmetry. On the other hand, the theory also has an anomalous chiral symmetry, the R-symmetry, under which we can take all of the scalar fields to

be neutral. So the theory is symmetric under this R-symmetry, combined with a simultaneous shift:

$$S \to S + i(N - N_f)\alpha. \tag{13.34}$$

Any superpotential must transform with charge 2 under this symmetry. The field Φ is neutral. But the Λ parameter transforms:

$$\Lambda = e^{-\frac{8\pi^2}{b_0 g^2}} = e^{-\frac{8\pi^2}{3N-N_f}S} \tag{13.35}$$

so

$$\Lambda^{\frac{3N-N_f}{N-N_f}} \to e^{2i\alpha}\Lambda^{\frac{3N-N_f}{N-N_f}}. \tag{13.36}$$

13.5 The superpotential in the case $N_f < N - 1$

Consider first the case $N_f < N - 1$. At energies well below the scale v, the theory consists of a pure (supersymmetric) $SU(N - N_f)$ gauge theory, and a number of neutral chiral multiplets. The chiral multiplets can couple to the gauge theory only through non-renormalizable operators. Because the moduli are neutral, there are no dimension-four couplings. There are possible dimension-five couplings; they are of the form

$$\delta\phi W_\alpha^2 \tag{13.37}$$

where $\delta\phi$ represents the fluctuations of the moduli fields about their expectation values; the coefficient of this operator will be of order $1/v$.

We can be more precise about the form of this coupling by noting that it must respect the various symmetries, if it is written in terms of the original, unshifted fields (this is similar to our argument for the form of the superpotential). In particular, a coupling of the form:

$$\mathcal{L}_{\text{coup}} = (S + a\ln(\Phi))W_\alpha^2 \tag{13.38}$$

respects all of the symmetries. It clearly respects the $SU(N_f)$ symmetries. It respects the non-anomalous $U(1)_R$ symmetry as well, for a suitable choice of a, since

$$\ln(\Phi) \to \ln(\Phi) + (N - N_f)/N_f\alpha. \tag{13.39}$$

It is not hard to see how this coupling is generated.

$$\Phi \approx v^N + v^{N-1}\phi. \tag{13.40}$$

Im ϕ couples to $F\tilde{F}$ through the anomaly diagram, just like an axion. The real part couples to F^2. One can see this by a direct calculation, or by noting that the masses of the heavy fields are proportional to v, so the gauge coupling of the $SU(N - N_f)$

theory depends on v:

$$\alpha_{N-N_f}^{-1}(\mu) = \alpha_N^{-1}(v) + \frac{b_0^{(N-N_f)}}{4\pi} \ln(\mu/v). \tag{13.41}$$

Since $\Phi \sim v^{N_f}$, we see that we have precisely the correct coupling. It is easy to see which Feynman graphs generate the couplings to the real and imaginary parts.

But we have seen that in the $SU(N - N_F)$ theory, gaugino condensation gives rise to a superpotential for the coefficient of W_α^2; in this case, this is precisely

$$W = \frac{\Lambda^{\frac{3N-N_f}{N-N_f}}}{\Phi^{\frac{1}{N-N_f}}}. \tag{13.42}$$

So we have understood the origin of the superpotential in these theories.

13.6 $N_f = N - 1$: the instanton-generated superpotential

In the case $N_f = N - 1$, the superpotential is generated by a different mechanism: instantons. Before describing the actual computation, we give some circumstantial evidence for this fact. Consider the instanton action. This is

$$e^{\frac{-8\pi^2}{g^2(v)}}. \tag{13.43}$$

Here we have assumed that the coupling is to be evaluated at the scale of the scalar vevs. The gauge group is, after all, completely broken, so provided that the computation is finite, this is the only relevant scale (we are also assuming that all of the vevs are of the same order). So any superpotential we might compute behaves as

$$W \sim v^3 \left(\frac{\Lambda}{v}\right)^{2N+1} \sim \frac{\Lambda^{2N+1}}{v^{2N-2}}, \tag{13.44}$$

which is the behavior predicted by the symmetry arguments.

To actually compute the instanton contribution to the superpotential, we need to develop further than in Chapter 5 the instanton computation and the structure of the supersymmetry zero modes. The required techniques were developed by 't Hooft, when he computed the baryon-number-violating terms in the effective action of the standard model; 't Hooft started by noting that, in the presence of the Higgs field, there *is no instanton solution.* This can be seen by a simple scaling argument. The instanton solution will now involve A^μ and ϕ. Suppose one has such a solution. Now simply do a rescaling of all lengths:

$$x^\mu \to \rho x^\mu; A^\mu \to \frac{1}{\rho} A^\mu; \quad \phi \to \phi \tag{13.45}$$

(because ϕ must tend to its expectation value at ∞, we cannot rescale it). Then the gauge kinetic terms are invariant, but the scalar kinetic terms are not; $|D\phi|^2 \to \rho^2 |D\phi|^2$. So the action is changed, and there is no solution.

However, the instanton configuration, while not a solution, is still distinguished by its topology; 't Hooft argued that it makes sense to integrate over solutions of a given topology. This just means that we write a configuration for each value of ρ, and integrate over ρ. For small ρ, we can understand this in the following way. The non-zero modes of the instanton, before turning on the scalar vevs, all have eigenvalues of order $1/\rho$ or larger, and can be ignored. There are also zero modes. Those associated with rotations and translations will remain at zero, even in the presence of the scalar, since they correspond to exact symmetries. But this is not the case for the dilatational zero mode; this mode is slightly lifted. The scaling argument above shows that the action is smallest at small ρ; we will see in a moment that the action of the interesting configurations vanishes as $\rho \to 0$. We know from our earlier studies of QCD, on the other hand, that the renormalization of the coupling tends to make the action large at small ρ. Together, these effects yield a minimum of the action at small, finite ρ, giving a self-consistent justification of the approximation.

To proceed with the computation, we review 't Hooft's notation for the instanton. Recall that the instanton preserves an $SU(2)$, which is a combination of rotations and gauge transformations. Since the instanton solution is a Euclidean solution, the original spatial symmetry is $O(4)$, so it is valuable to understand how the rotational $SU(2)$ is embedded in this group. More precisely, $SU(2) \times SU(2) \subset O(4)$. Our two-component notation is convenient for this. After the rotation to Euclidean space, we can write the matrices σ^μ and $\bar{\sigma}^\mu$ as

$$\sigma^\mu = (i, \vec{\sigma}) \quad \bar{\sigma}^\mu = (i, -\vec{\sigma}). \tag{13.46}$$

The $O(4)$ generators are:

$$\begin{array}{ll} \sigma^{4i} = \sigma^i & \sigma^{ij} = \epsilon_{ijk}\sigma^k \\ \bar{\sigma}^{4i} = -\sigma^i & \sigma^{ij} = \epsilon_{ijk}\sigma^k. \end{array} \tag{13.47}$$

Note we have separated out the indices 4 and i. The matrices $\sigma^{\mu\nu}$ and $\bar{\sigma}^{\mu\nu}$ are self-dual and anti-self-dual respectively, as can be seen from these equations.

Then 't Hooft introduced a symbol, $\eta_{a\mu\nu} = 1/2\mathrm{Tr}\,(\sigma^a \sigma^{\mu\nu})$, and a similar symbol $\bar{\eta}$. These symbols are self-dual and anti-self-dual, and define an embedding of the two $SU(2)$ subgroups. The detailed components are:

$$\eta_{a\mu\nu} = \epsilon_{a\mu\nu} \text{ if } \mu, \nu = 1, 2, 3 \quad \eta_{a4\nu} = -\delta_{a\nu} \quad \eta_{a\mu 4} = \delta_{a\mu} \quad \eta_{a44} = 0. \tag{13.48}$$

Note that $\bar{\eta}$ differs by:

$$\bar{\eta}_{a\mu\nu} = (-1)^{\delta_{\mu 4} + \delta_{\nu 4}} \eta_{a\mu\nu} \tag{13.49}$$

Also, η has the properties:

$$\eta_{a\mu\nu} = -\eta_{av\mu} \quad \eta_{a\mu\nu} = \frac{1}{2}\epsilon_{\mu\nu\alpha\beta}\eta_{a\alpha\beta} \tag{13.50}$$

In terms of η, the instanton solution is:

$$A^a_\mu(x) = \frac{2\eta_{a\mu\nu}x_\nu}{(x^2+\rho^2)}. \tag{13.51}$$

It is straightforward to work out $F_{\mu\nu}$ (see the exercises):

$$F^a_{\mu\nu} = \frac{\eta_{a\mu\nu}}{(x^2+\rho^2)^2}. \tag{13.52}$$

F is self-dual, since η is, so this is a solution of the Euclidean equations.

A second-rank antisymmetric tensor ($F_{\mu\nu}$) is a six-dimensional representation of $SO(4)$; under $SU(2) \times SU(2)$, it decomposes as a $(3, 1) + (1, 3)$, where these are the self-dual and anti-self-dual parts of the tensor. The η symbol is essentially a Clebsch–Gordan coefficient, which describes a mapping of one $SU(2)$ subgroup of $SO(4)$ into $SU(2)$. Other important properties of these symbols include

$$\epsilon_{abc}\eta_{b\mu\nu}\eta_{c\kappa\lambda} = \delta_{\mu\kappa}\eta_{av\lambda} - \delta_{\mu\lambda}\eta_{av\kappa} - \delta_{\nu\kappa}\eta_{a\mu\lambda} + \delta_{\nu\lambda}\eta_{a\mu\kappa}. \tag{13.53}$$

At large distances, the instanton is a gauge transformation of "nothing." The gauge transformation is just

$$g^i_j = i\bar{\sigma}^{\mu\ i}_{\ j}x^\mu. \tag{13.54}$$

This can be thought of as a mapping of S_3 into $SU(2)$; the winding number of the instanton just counts the number of times space is mapped onto the group.

In this form, it is useful to note another way to describe the instanton solution. By an inversion of coordinates, one can write:

$$A^a_\mu = \frac{2}{g^2}\frac{\rho^2}{(x^2+\rho^2)}\eta_{a\mu\nu}\frac{x^\nu}{x^2}. \tag{13.55}$$

This *singular gauge* instanton is often useful since it falls off more rapidly at large x than the original instanton solution.

Now for the doublets we solve the equation:

$$D^2Q = D^2\bar{Q} = 0. \tag{13.56}$$

This has the solutions

$$Q^i = \bar{Q}^{i\ \dagger} = i\bar{\sigma}^{\mu i}_j x^\mu \left(\frac{1}{x^2+\rho^2}\right)^{1/2}\langle Q^j\rangle, \tag{13.57}$$

and similarly for $\bar{Q}$. Like the solution for A^μ, these solutions are "pure gauge" configurations as $r \to \infty$, i.e. they are gauge transformations by g of the constant vev. (Note here and above, the σ^μs are the Euclidean versions of the two-component Dirac matrices, $\sigma^\mu = (i, \vec{\sigma})$, $\bar{\sigma}^\mu = (i, -\vec{\sigma})$.)

The action of this configuration is:

$$S(\rho) = \frac{1}{g^2}(8\pi^2 + 4\pi^2\rho^2 v^2). \tag{13.58}$$

Some features of this result are worth noting.

(1) The integral over ρ now converges for large ρ, since it is exponentially damped.
(2) Terms in the potential, involving $|Q|^4$, make smaller contributions to the action, by powers of ρ. Rescaling $x \to \rho x$, one sees that these terms are of order ρ^4. But ρ is at most of order $gv^{-1} = m_w$ (item (1) above), so these terms are suppressed. This justifies our neglect of these terms in the equations of motion.

Our goal is to compute the instanton contribution to the effective action. We particularly want to see if the instanton generates the conjectured non-perturbative superpotential. In order to compute the effective action, we need to ask about the fermion zero modes. Before turning on the vevs for the scalars, there are six zero modes. Two of these are generated by supersymmetry transformations of the instanton solution:

$$\delta\lambda = \sigma_\alpha^{\mu\nu\ \beta} F^{\mu\nu}\epsilon_\beta \tag{13.59}$$

so

$$\lambda^{\mathrm{SS}\ [\beta]}_{\alpha\ a} = \frac{8\sigma_\alpha^{\mu\ a\ \beta}}{(x^2+\rho^2)^2}. \tag{13.60}$$

Note that, because of the anti-self-duality of $\bar{\sigma}^{\mu\nu}$, two of the supersymmetry generators annihilate the lowest-order solution, i.e. there are only two supersymmetry zero modes. If we neglect the Higgs, the classical Yang–Mills action has a conformal (scale) symmetry. This is the origin of the ρ zero mode in the classical solution. In the supersymmetric case, there is, apart from supersymmetry, an additional fermionic symmetry called superconformal invariance. In superspace, the corresponding generators are

$$Q^{\mathrm{SC}} = \not{x} Q \tag{13.61}$$

so

$$\lambda^{\mathrm{SC}\ [\beta]}_{\alpha\ a} = \frac{8\ \not{x}\sigma_\alpha^{\mu\ a\ \beta}}{(x^2+\rho^2)^2}. \tag{13.62}$$

There are also two matter-field zero modes, one for each of the quark doublets:

$$\psi^i_{Q\,\alpha} = \frac{\delta^i_\alpha}{(x^2+\rho^2)^{3/2}} = \psi_{\bar{Q}} \tag{13.63}$$

(in the last equation we are treating $\bar{Q}$ also as a doublet; one can treat this as a 2^* representation by multiplying by ϵ_{ij}).

When we turn on the scalar vevs, these modes are corrected. The superconformal symmetry is broken by these vevs and, not surprisingly, the superconformal zero modes are lifted. In fact, they pair with the two quark zero modes. We can compute this pairing by treating the Yukawa terms in the Lagrangian as a perturbation, replacing the scalar fields by their classical values. Expanding to second order, i.e. including

$$\int d^4x\, Q^* \psi_Q \lambda \int d^4x'\, \bar{Q}^* \psi_{\bar{Q}} \lambda \tag{13.64}$$

and expanding the fields in the lowest-order eigenmodes, the superconformal and matter-field zero modes can be absorbed by these terms. Note, in particular, that both $Q_{\rm cl}$ and $\lambda^{\rm SC}$ are odd under $x \to -x$, while the matter-field zero modes are even, so the integral is non-zero. The supersymmetry zero modes, being even, cannot be soaked up in this way.

The wave functions of the supersymmetry zero modes are altered in the presence of the Higgs fields, and they now have components in the ψ^*_Q and $\psi^*_{\bar{Q}}$ directions. For ψ_Q, for example, we need to solve the equation

$$D_\mu \bar{\sigma}^\mu \psi^{\rm SS*}_Q = \lambda^{\rm SS} Q^*. \tag{13.65}$$

But this equation is easy to solve, starting with our solution of the scalar equation. If we simply take:

$$\psi^{\rm SS}_Q = D_\mu \sigma^\mu Q^*, \tag{13.66}$$

plugging back in, the left-hand side becomes

$$D^2 Q + \sigma_{\mu\nu} F^{\mu\nu} Q, \tag{13.67}$$

but the first term vanishes on the classical solution, while the second is, indeed, just $\lambda^{\rm SS} Q^*$.

With these ingredients, we can compute the superpotential terms in the effective action. In particular, the non-perturbative superpotential predicts a non-zero term in the component form of the effective action proportional to:

$$\frac{\partial^2 W}{\partial Q \partial \bar{Q}} = \frac{1}{v^4} \psi_Q \psi_{\bar{Q}}. \tag{13.68}$$

We can calculate this term by studying the corresponding Green function. We need to be careful, now, about the various collective coordinates. We want to study the gauge-invariant correlation function

$$\langle \bar{Q}(x)\psi_Q(x)\psi_{\bar{Q}}(y)Q(y)\rangle \tag{13.69}$$

in the presence of the instanton. Since we are interested in the low-momentum limit of the effective action, we can take x and y to be widely separated. We need to integrate over the instanton location, x_0, and the instanton orientation and scale size. Because the gauge fields are massive, we can take x and y both to be far away from the instanton. Then, from our explicit solution for the supersymmetry zero modes

$$\psi_Q(x) \propto \not{D} Q \propto \not{D} \frac{i\sigma^\mu (x^\mu - x_0^\mu)}{((x-x_0)^2+\rho^2)^{1/2}} \to g(x-x_0)S_F(x-x_0), \tag{13.70}$$

with a similar equation for $\psi_{\bar{Q}}$. The g and $g^\dagger$ factors are canceled by corresponding factors in Q and $\bar{Q}$, at large distances. Substituting these expressions into the path integral and integrating over x_0 gives a convolution, $v^2 \int d^4x_0 S_F(x-x_0)S_F(y-y_0)$. Extracting the external propagators, we obtain the effective action. Integrating over ρ gives a term of precisely the desired form. If we contract the gauge and spinor indices in a gauge- and rotationally invariant manner, the integral over rotations just gives a constant factor. It is some work to do all of the bookkeeping correctly. The evaluation of the determinant is greatly facilitated by supersymmetry: there is a precise fermion–boson pairing of all of the non-zero modes. In the exercises, you are asked to work out more details of this computation; further details can also be found in the References.

Without working through all of the details we can see the main features.

(1) The perturbative lifting of the zero modes gives rise to a contribution proportional to v^2 (see Fig. 13.1).
(2) The matter-field component of the supersymmetry zero modes studied above gives a contribution to the gauge-invariant correlation function:

$$v^4 \int d^4x_0 S_f(x-x_0)S_f(y-y_0) \tag{13.71}$$

(3) The integral over the gauge collective coordinates (equivalently the rotational collective coordinates) simply gives a constant, since we have computed a gauge- and rotationally invariant quantity.
(4) The scale size collective coordinate integral behaves as

$$W = A \int d\rho v^4 e^{-\left(\frac{8\pi^2}{g^2(\rho)} + 4\pi^2\rho^2 v^2\right)} \tag{13.72}$$

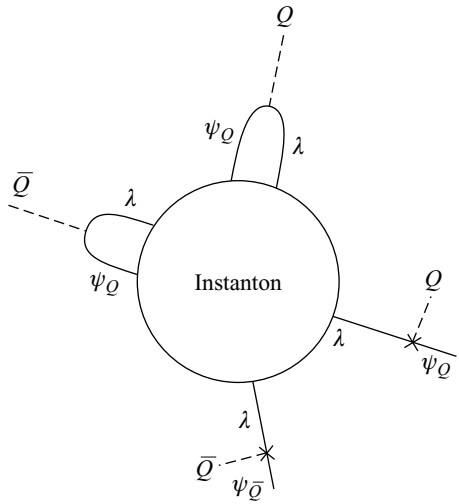

Fig. 13.1. Schematic description of the instanton computation of the superpotential. Four zero modes are tied together by the scalar vevs; two gluino zero modes turn into ψ zero modes as well.

where the power of ρ has been determined from dimensional analysis, and A is a constant.

(5) Extracting the constant requires careful attention to the normalization of the zero modes and to the Jacobians for the collective coordinates. However, the non-zero modes come in Fermi–Bose pairs, and their contribution to the functional integral cancels.

(6) The final ρ integral gives

$$W = A' \frac{\Lambda^5}{v^2} \tag{13.73}$$

consistent with the expectations of the symmetry analysis.

This analysis generalizes straightforwardly to the case of general N_c.

13.6.1 An application of the instanton result: gaugino condensation

The instanton calculation for the case $N_f = N - 1$ is a systematic, weak-coupling computation of the superpotential which appears in the low-energy-effective action. Seiberg has noted that this result, plus holomorphy, allows systematic study of the strongly coupled regime of other theories. To understand this, take $N = 2$, and add a mass term for the quark. In this case, for very small mass, the superpotential is:

$$W = m\bar{Q}Q + \frac{\Lambda^{2N+1}}{\bar{Q}Q}. \tag{13.74}$$

We can solve the equation for Q,

$$Q_1 = \begin{pmatrix} 0 \\ v \end{pmatrix} \quad v = \left(\frac{\Lambda^5}{m}\right)^{1/4}. \tag{13.75}$$

Using this, we can evaluate the expectation value of the superpotential at the minimum:

$$W(m, \Lambda) = \Lambda^{5/2} m^{1/2}. \tag{13.76}$$

Because W is holomorphic, this result also holds for large m. For large m, the low-energy theory is just a pure $SU(2)$ gauge theory. We expect, there, that the superpotential is $\langle\lambda\lambda\rangle = \Lambda_{\rm le}^3$. But this is equal to:

$$W = \langle\lambda\lambda\rangle = m^3 e^{-\frac{8\pi^2}{2g^2(m)}}. \tag{13.77}$$

The right-hand side is nothing but $\Lambda_{\rm le}^3$. We have, in fact, done a systematic, reliable computation of the gluino condensate in a strongly interacting gauge theory!

Suggested reading

Excellent treatments of supersymmetric dynamics appear in the text by Weinberg (1995), and in Michael Peskin's lectures (1997). We have already mentioned 't Hooft's original instanton paper (1976). The instanton computation of the superpotential is described in Affleck *et al*. (1984).

Exercises

(1) Verify that $\sigma_{\mu\nu}$ and $\bar{\sigma}_{\mu\nu}$ are self-dual and anti-self-dual, respectively. This means $\mathrm{Tr}\sigma^a\sigma_{\mu\nu}$ is a self-dual tensor. Verify the connection to η; do the same thing for $\bar{\eta}$.

(2) Verify Eq. (13.52), so F is self-dual, and so solves the Euclidean Yang–Mills equations. Check that asymptotically the instanton potential is a gauge transform of "nothing."

(3) Verify the solution of the scalar field equation (Eq. (13.57)). Compute the action of this field configuration.

(4) Do the zero-mode counting for the case of general $N_{\rm c}$, $N_{\rm f} = N_{\rm c} - 1$. Show that, again, all but two zero modes pair with matter-field zero modes; two supersymmetry zero modes contain matter-field components which can give rise to the expected superpotential.

14

Dynamical supersymmetry breaking

One of the original reasons for interest in supersymmetry was the possibility of dynamical supersymmetry breaking. So far, however, we have exhibited models in which supersymmetry is unbroken, as in the case of QCD with only massive quarks, or models with moduli spaces or approximate moduli spaces. In this section, we describe a number of models in which non-trivial dynamics breaks supersymmetry. We will see that the dynamical supersymmetry breaking occurs under special, but readily understood, conditions. In some cases, we will be able to exhibit this breaking explicitly, through systematic calculations. In others, we will have to invoke more general arguments.

14.1 Models of dynamical supersymmetry breaking

We might ask why, so far, we have not found supersymmetry to be dynamically broken. In supersymmetric QCD with massive quarks, we might give the index as an explanation. We might also note that there is not a promising candidate for a goldstino. With massless quarks, we have flat directions, and as the fields get larger, the theory becomes more weakly coupled, so any potential tends to zero.

This suggests two criteria for finding models with dynamical supersymmetry breaking.

(1) The theory should have no flat directions at the classical level.
(2) The theory should have a spontaneously broken global symmetry.

The second criterion implies the existence of a Goldstone boson. If supersymmetry is unbroken, any would-be Goldstone boson must lie in a multiplet with another scalar, as well as a Weyl fermion. This other scalar, like the Goldstone particle, has no potential, so the theory has a flat direction. But by assumption, the theory classically (and therefore almost certainly quantum mechanically) has no flat direction.

So supersymmetry is likely to be broken. These criteria are heuristic but, in practice, when a systematic analysis is possible, they are always correct.

Perhaps the simplest model with these features is a supersymmetric $SU(5)$ theory with a single $\bar{5}$ and 10. In the Exercises, you will show that this theory, in fact, has no flat directions, and that it has two non-anomalous $U(1)$ symmetries. One can give arguments that these symmetries are broken. So it is likely that this theory breaks supersymmetry.

However, this is a strongly coupled model, and it is difficult to prove that supersymmetry is broken. In the next section, we will describe a simple weakly coupled theory in which dynamical supersymmetry breaking occurs within a controlled approximation.

14.1.1 The $(3, 2)$ *model*

A model in which supersymmetry turns out to be broken is the "3–2 model." This theory has gauge symmetry $SU(3) \times SU(2)$, and matter content:

$$Q(3, 2) \quad \bar{U}(\bar{3}, 1) \quad L(1, 2) \quad \bar{D}(\bar{3}, 1). \tag{14.1}$$

This is similar to the field content of a single generation of the standard model, without the extra $U(1)$ and the positron. The most general renormalizable superpotential consistent with the symmetries is

$$W = \lambda Q L \bar{U}. \tag{14.2}$$

This model admits an R-symmetry that is free of anomalies. There is also a conventional $U(1)$ symmetry, under which the charges of the various fields are the same as in the standard model (one can gauge this symmetry if one also adds an e^+ field).

While this model has global symmetries, it is different from supersymmetric QCD in that it does not have classical flat directions. To see this, note that by $SU(3) \times SU(2)$ transformations, one can bring Q to the form

$$Q = \begin{pmatrix} a & 0 \\ 0 & b \\ 0 & 0 \end{pmatrix}. \tag{14.3}$$

Now the vanishing of the $SU(2)$ D term forces

$$L = \left(0, \sqrt{|a^2| - |b^2|}\right). \tag{14.4}$$

The vanishing of the F terms for $\bar{u}$ require $|a| = |b|$. Then the vanishing of the $SU(3)$ D term forces

$$\bar{U} = \begin{pmatrix} a' \\ 0 \\ 0 \end{pmatrix} \quad \bar{D} = \begin{pmatrix} 0 \\ a'' \\ 0 \end{pmatrix} \tag{14.5}$$

(up to interchange of the two vevs), with

$$|a'| = |a''| = |a|.$$

Finally, the $\partial W/\partial L$ equations lead to $a = 0$.

To analyze the dynamics of this theory, consider first the case that $\Lambda_3 \gg \Lambda_2$. Ignoring, at first, the superpotential term, this is just $SU(3)$ with two flavors. In the flat direction of the D terms, there is a non-perturbative superpotential,

$$W_{\rm np} = \frac{\Lambda^5}{\det \overline{Q} Q} \sim \frac{1}{v^4}. \tag{14.6}$$

The full superpotential in the low-energy theory is a sum of this term and the perturbative term. It is straightforward to minimize the potential, and establish that supersymmetry is broken. One finds

$$a = 1.287\Lambda/\lambda^{1/7} \quad b = 1.249\Lambda/\lambda^{1/7} \quad E = 3.593\lambda^{10/7}\Lambda^4. \tag{14.7}$$

If $\Lambda_2 \gg \Lambda_3$, supersymmetry is still broken, but the mechanism is different. In this case, before including the classical superpotential, the strongly coupled theory is $SU(2)$ with two flavors. This is an example of a model with a *quantum moduli space*. This notion will be explained in the next chapter, but it implies that $\langle QL \rangle \neq 0$, so at low energies there is a superpotential (F-term) for $\bar{U}$.

There does not, at the present time, exist an algorithm to generate all models which exhibit dynamical supersymmetry breaking, but many classes are known. A generalization of the $SU(5)$ model, for example, is provided by an $SU(N)$ model with an antisymmetric tensor field, A_{ij}, and $N - 4$ $\bar{F}$s. It is also necessary to include a superpotential,

$$W = \lambda_{ab} A \bar{F}^a \bar{F}^b. \tag{14.8}$$

Other broad classes are known, including generalizations of the $(3, 2)$ model. A somewhat different, and particularly interesting set of models, is described in Section 15.4. Catalogs of known models, as well as studies of their dynamics, are given in some of the references in the Suggested reading at the end of this chapter.

We have seen, in this section, that dynamical breaking of supersymmetry is common. Flat directions are often lifted, and in many instances, supersymmetry is broken with a stable ground state. So we are ready to address the question: how might supersymmetry be broken in the real world?

14.2 Particle physics and dynamical supersymmetry breaking

14.2.1 Gravity mediation and dynamical supersymmetry breaking: anomaly mediation

One simple approach to model building which we explored in Chapter 11 was to treat a theory which breaks supersymmetry as a "hidden sector." This construction,

as we presented it, was rather artificial. If we replace, say, the Polonyi sector, by a sector which breaks supersymmetry dynamically, the situation is dramatically improved. If we suppose that there are some fields transforming under only the Standard Model gauge group, and some transforming under only the gauge group responsible for symmetry breaking, the visible/hidden sector division is automatic. As we will see, this sort of division can arise rather naturally in string theory.

In such an approach the scale of supersymmetry breaking is again $m_{3/2}M_{\rm p}$, where we now understand this scale as arising as the exponential of a small coupling at a high-energy scale (presumably the Planck, GUT, or string scale). For scalars, soft-supersymmetry-breaking masses and couplings arise just as they did previously. There is no symmetry reason why these masses should exhibit any sort of universality.

One puzzle in this scenario is related to gluino masses. Examining the supergravity Lagrangian, the only terms which can lead to gaugino masses are

$$\mathcal{L}_{\lambda\lambda} = e^{K/2} f'_{\alpha\beta k}(D_k W)\lambda^\alpha \lambda^\beta. \tag{14.9}$$

Here f is the gauge coupling function. So in order to obtain a substantial gaugino mass, it is necessary that there be gauge-single fields with non-zero F terms. In most models of dynamical supersymmetry breaking, there are no scalars which are singlets under all of the gauge interactions. Even when there are, it is necessary to suppose that there is some sort of discrete symmetry which accounts for the absence of certain couplings. These symmetries will forbid coupling of hidden sector fields to visible sector gauge fields through low-dimension operators. In other words, we don't have couplings of the form

$$\frac{S}{M} W_\alpha^2 \tag{14.10}$$

where the F component of S has a non-zero vev. This suggests that gaugino masses would be suppressed relative to squark and slepton masses by powers of $M_{\rm int}/M_{\rm p}$.

But this turns out not to be quite correct; gaugino masses, in supergravity theories, can arise as a result of certain anomalous behavior. Suppose that the low-energy gauge group is just that of the Standard Model, and that there is some very massive field, Σ, say an octet of $SU(3)$. The Σ superpotential might take the form:

$$W_\Sigma = \frac{M_\Sigma}{2} \Sigma^a \Sigma^a. \tag{14.11}$$

As a result of supersymmetry breaking, there will in general be a term in the *potential* for the Σ scalars:

$$V_B(\Sigma) = b\, m_{3/2} M_\Sigma \Sigma^2 + \text{c.c.} \tag{14.12}$$

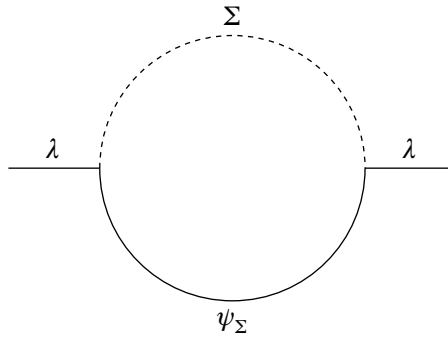

Fig. 14.1. Feynman diagram associated with anomaly mediation.

As a result, there is a one-loop contribution to the *gluino* masses upon integrating out the Σ field, as in Fig. 14.1. This contribution is of order $b(\alpha_s/\pi)m_{3/2}$. But we have just argued that there can be no such term from a supersymmetric effective action.

There is a simple resolution to this paradox. The puzzling contribution to the gaugino mass is independent of M. Suppose, then, we "regulated" this finite diagram with a Pauli–Villars regulator field. Then the gaugino mass would vanish, as expected from symmetry arguments, and the paradox would seem to disappear.

But now we have a different problem. Suppose, for a moment, that the light squark fields were also vector-like. Then we might expect that we would have to regulate these with Pauli–Villars fields as well. But then, these Pauli–Villars fields *would* give contributions to gaugino masses, of order $(\alpha_s/\pi)m_{3/2}$. In fact, we would obtain:

$$m_{\lambda_i} = b_0 \frac{\alpha_i}{\pi} m_{3/2}, \tag{14.13}$$

where b_0 is the one-loop beta function. This mechanism for generating gaugino masses is known as *anomaly mediation*.

In the framework we have described, there is no reason to expect that the leading contribution to scalar masses should vanish. But one might speculate that this could occur in some circumstances. If it does, and if the one-loop contributions to scalar masses vanish as well, there are two-loop contributions, analogous to those we have found above. They can be obtained by a similar analysis. The result is proportional to two powers of the couplings, α_i, and the beta functions. At one level, this formula is quite appealing. It appears highly predictive. It is also universal, so flavor changing processes are suppressed. It is also, unfortunately, phenomenologically unsuccessful, as it predicts tachyonic masses for slepton doublets.

To build models along these lines, then, two ingredients are required. First, one must explain why there are almost no contributions to scalar masses as large as these

two-loop effects. Then one must explain why there are some which can give larger masses to the lepton doublets. The first turns out to be problematic. In string theory and in certain large extra dimension models, the leading-order Kahler potential is often of a special form which can give vanishing scalar masses even in the presence of hidden sector supersymmetry breaking. But in these situations, there are corrections to the Kahler potential which usually swamp the anomalous two-loop contributions.

14.2.2 Low-energy dynamical supersymmetry breaking: gauge mediation

An alternative to the conventional supergravity approach is to suppose that supersymmetry is broken at some much-lower energy, with gauge interactions serving as the messengers of supersymmetry breaking. The basic idea is simple. One again supposes that one has some set of new fields and interactions which break supersymmetry. Some of these fields are taken to carry ordinary Standard Model quantum numbers, so that "ordinary" squarks, sleptons and gauginos can couple to them through gauge loops. This approach, which is referred to as "gauge mediated supersymmetry breaking" (GMSB), has a number of virtues.

(1) It is highly predictive: as few as two parameters describe all soft breakings.
(2) The degeneracies required to suppress flavor-changing neutral currents are automatic.
(3) GMSB easily incorporates DSB, and so can readily explain the hierarchy.
(4) GMSB makes dramatic and distinctive experimental predictions.

The approach, however, also has drawbacks. Perhaps most serious is related to the "μ problem," which we discussed in the MSSM. In theories with high-scale-supersymmetry breaking, we saw that there is not really a problem at all; a μ term of order the weak scale is quite natural. The μ problem, however, finds a home in the framework of low-energy breaking. The difficulty is that, if one is trying to explain the weak scale dynamically, one does not want to introduce the μ term by hand. Various solutions have been offered for this problem, but none is yet compelling. In the rest of our discussion, we will simply assume that a μ term has been generated in the effective theory, and not worry about its origin.

Minimal Gauge Mediation (MGM)

The simplest model of gauge mediation contains, as messengers, a vector-like set of quarks and leptons, q, $\bar{q}$, ℓ and $\bar{\ell}$. These have the quantum numbers of a 5 and $\bar{5}$ of $SU(5)$. The superpotential is taken to be

$$W_{\rm mgm} = \lambda_1 q\bar{q} + \lambda_2 S\ell\bar{\ell}. \tag{14.14}$$

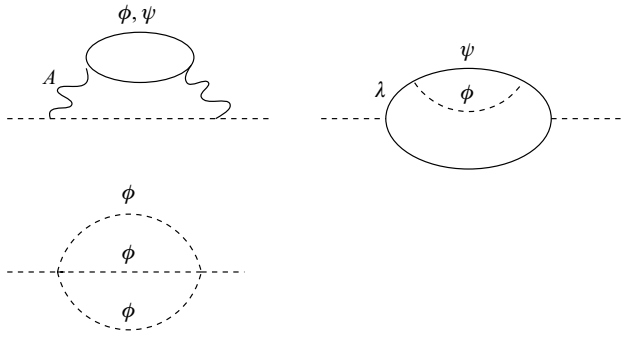

Fig. 14.2. Two-loop diagrams contributing to squark masses in a simple model of gauge mediation.

We suppose that some dynamics gives rise to non-zero expectation values for S and F_S. Here we won't provide a complete microscopic model, which explains the origin of the parameters F_S and $\langle S \rangle$ that will figure in our subsequent analysis; some constructions are described in the suggested reading. It is a good research problem to find a compelling model of the underlying dynamics. Instead, we will go ahead and immediately compute the superparticle spectrum for such a model. Ordinary squarks and sleptons gain mass through the two-loop diagrams shown in Fig. 14.2. While the prospect of computing a set of two-loop diagrams may seem intimidating, the computation is actually quite easy. If one treats F_S/S as small, there is only one scale in the integrals. It is a straightforward matter to write down the diagrams, introduce Feynman parameters, and perform the calculation. There are also various non-trivial checks. For example, the sum of the diagrams must vanish in the supersymmetric limit.

One obtains the following expressions for the scalar masses:

$$\widetilde{m}^2 = 2\Lambda^2 \left[C_3 \left(\frac{\alpha_3}{4\pi}\right)^2 + C_2 \left(\frac{\alpha_2}{4\pi}\right)^2 + \frac{5}{3}\left(\frac{Y}{2}\right)^2 \left(\frac{\alpha_1}{4\pi}\right)^2 \right], \tag{14.15}$$

where $\Lambda = F_S/S$, and $C_3 = 4/3$ for color triplets and zero for singlets, $C_2 = 3/4$ for weak doublets and zero for singlets. For the gaugino masses one obtains:

$$m_{\lambda_i} = C_i \, \frac{\alpha_i}{4\pi} \, \Lambda. \tag{14.16}$$

This expression is valid only to lowest order in Λ. Higher-order corrections have been computed; it is straightforward to compute exactly in Λ.

All of these masses are positive, and they are described in terms of a single new parameter, Λ. The lightest new particles are the partners of the $SU(3) \times SU(2)$ singlet leptons. If their masses are of order 100 GeV, we have that $\Lambda \sim 30$ TeV. The spectrum has a high degree of degeneracy. In this approximation, the masses

of the squarks and sleptons are functions only of their gauge quantum numbers, so flavor-changing processes are suppressed. Flavor violation arises only through Yukawa couplings, and these can appear only in graphs at high loop order. It is further suppressed because all but the top Yukawa coupling is small.

Apart from the parameter Λ, one has the μ and $B\mu$ parameters (both complex), for a total of five. This is three beyond the minimal Standard Model. If the underlying susy-breaking theory conserves CP, this can eliminate the phases, reducing the number of parameters by two.

$SU(2) \times U(1)$ breaking

At lowest order, all of the squark and slepton masses are positive. The large top quark Yukawa coupling leads to large corrections to $m_{H_U}^2$, however, which drive $SU(2) \times U(1)$ breaking. The calculation is just a repeat of one we have done in the case of the MSSM. Treating the mass of $\tilde{t}$ as independent of momentum is consistent, provided we cut the integral off at a scale of order Λ (at this scale, the calculation leading to Eq. (14.15) breaks down, and the propagator falls rapidly with momentum) and we have

$$m_{H_U}^2 = \left(m_{H_U}^2\right)_0 - \frac{6y_t^2}{16\pi^2} \ln\left(\Lambda^2/\widetilde{m}_t^2\right)\left(\widetilde{m}_t^2\right)_0. \tag{14.17}$$

While the loop correction is nominally a three-loop effect, because the stop mass arises from gluon loops while the Higgs mass arises at lowest order from W loops, we have a substantial effect

$$\left(\frac{\widetilde{m}_t^2}{m_{H_U}^2}\right)_0 = \frac{16}{9}\left(\frac{\alpha_3}{\alpha_2}\right)^2 \sim 20 \tag{14.18}$$

and the Higgs mass-squared is negative. These contributions are quite large, and it is usually necessary to tune the μ-term and other possible contributions to the Higgs mass to obtain sufficiently small W and Z masses.

Light gravitino phenomenology

There are other striking features of these models. One of the more interesting is that the lightest supersymmetric particle, or LSP, is the gravitino. Its mass is

$$m_{3/2} = 2.5\left(\frac{F}{(100\ \text{TeV})^2}\right)\text{eV}. \tag{14.19}$$

The next-to-lightest supersymmetric particle, or NLSP, can be a neutralino, or a charged right-handed slepton. The NLSP will decay to its superpartner plus a gravitino in a time long compared with typical microscopic times, but still quite short. The lifetime can be determined from low-energy theorems, in a manner reminiscent of the calculation of the pion lifetime. Just as the chiral currents are

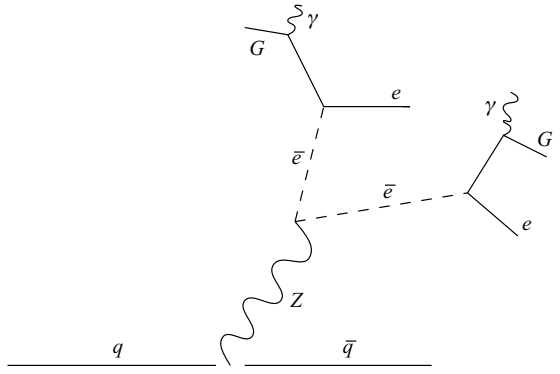

Fig. 14.3. Decay leading to $e^+e^-\gamma\gamma$ events.

linear in the (nearly massless) pion field,

$$j^{\mu 5} = f_\pi \partial^\mu \pi \quad \partial_\mu j^{\mu 5} = \partial^2 \pi \approx 0, \tag{14.20}$$

so the supersymmetry current is linear in the goldstino, G:

$$j^\mu_\alpha = F\gamma^\mu G + \sigma^{\mu\nu}\lambda F_{\mu\nu} + \cdots, \tag{14.21}$$

where F, here, is the goldstino decay constant. From this, if one assumes that the LSP is mostly photino, one can calculate the amplitude for $\tilde{\gamma} \to G + \gamma$ in much the same way one considers processes in current algebra. From Eq. (14.21), one sees that $\partial_\mu j^\mu_\alpha$ is an interpolating field for G, so:

$$\langle G\gamma|\tilde{\gamma}\rangle = \tfrac{1}{F}\langle\gamma|\partial_\mu j^\mu_\alpha|\tilde{\gamma}\rangle. \tag{14.22}$$

The matrix element can be evaluated by examining the second term in the current, Eq. (14.21), and noting that $\partial\!\!\!/\lambda = m_\lambda \lambda$.

Given the matrix element, the calculation of the NLSP lifetime is straightforward, and yields

$$\Gamma(\tilde{\gamma} \to G\gamma) = \frac{\cos^2\theta_W m^5_{\tilde{\gamma}}}{16\pi F^2}. \tag{14.23}$$

This yields a decay *length*:

$$c\tau = 130\left(\frac{100\text{ GeV}}{m_{\tilde{\gamma}}}\right)^5\left(\frac{\sqrt{F}}{100\text{ TeV}}\right)^4 \mu\text{m}. \tag{14.24}$$

In other words, if F is not too large, the NLSP may decay in the detector. One even has the possibility of measurable displaced vertices. The signatures of such low decay constants would be quite spectacular. Assuming the photino (bino) is the NLSP, one has processes such as $e^+e^- \to \gamma\gamma + \not{E}_t$ and $p\bar{b} \to e^+e^-\gamma\gamma + \not{E}_t$, as indicated in Fig. 14.3, where $\not{E}_t$ is the missing transverse energy.

Suggested reading

There are a number of good reviews of dynamical supersymmetry breaking, including those of Shadmi and Shirman (2000) and Terning (2003). The former includes catalogs of models and mechanisms. There is a large literature on gauge-mediated models and their phenomenology; a good review is provided by Giudice and Rattazzi (1999). A clear exposition of the origin of anomaly mediation is provided by Bagger *et al*. (2000) and in Weinberg's text (1995).

Exercises

(1) Check that the $SU(N)$ models, with an antisymmetric tensor and $N-4$ anti-fundamentals, have no flat directions, and that they have a non-anomalous $U(1)$ symmetry.
(2) Verify Eq. (14.13) using the Pauli–Villars procedure suggested in this chapter.

15

Theories with more than four conserved supercharges

In theories with more than four conserved supercharges (extended supersymmetry), the supersymmetry generators obey relations:

$$\left\{Q_\alpha^I, Q_{\dot\alpha}^J\right\} = \not{p}\delta^{IJ} \quad \left\{Q_\alpha^I, Q_\beta^J\right\} = Z^{IJ}\epsilon_{\alpha,\beta}. \tag{15.1}$$

The quantities Z^{IJ} are known as central charges. We will see that these can arise in a number of physically interesting ways.

In theories with four supersymmetries, we have seen that supersymmetry provides powerful constraints on the possible dynamics. Theories with more than four supercharges ($N > 1$ in four dimensions) are not plausible as models of the real world, but they do have a number of remarkable features. As in some of our $N = 1$ examples, these theories typically have exact moduli spaces. Gauge theories with $N = 4$ supersymmetry exhibit an exact duality between electricity and magnetism. Theories with $N = 2$ supersymmetry have a rich – and tractable – dynamics, closely related to important problems in mathematics. In all of these cases, supersymmetry provides remarkable control over the dynamics, allowing one to address questions which are inaccessible in theories without supersymmetry. Supersymmetric theories in higher dimensions generally have more than four supersymmetries, and a number of the features of the theories we study in this chapter will reappear when we come to higher-dimensional field theories and string theory.

15.1 $N = 2$ theories: exact moduli spaces

Theories with $N = 1$ supersymmetry are tightly constrained, but theories with more supersymmetry are even more highly constrained. We have seen that often in perturbation theory $N = 1$ theories have moduli; non-perturbatively, sometimes, these moduli are lifted. In theories with $N = 1$ supersymmetry, detailed analysis is usually required to determine whether the moduli acquire potentials at the quantum level. For theories with more supersymmetries ($N > 1$ in four dimensions; $N \geq 1$

in five or more dimensions), one can show rather easily that the moduli space is exact. Here we consider the case of $N = 2$ supersymmetry in four dimensions. These theories can also be described by a superspace, in this case built from two Grassmann spinors, θ and $\tilde{\theta}$. There are two basic types of superfields, called vector and hypermultiplets. The vectors are chiral with respect to both D_α and $\tilde{D}_\alpha$, and have an expansion, in the case of a $U(1)$ field:

$$\psi = \phi + \tilde{\theta}^\alpha W_\alpha + \tilde{\theta}^2 \bar{D}^2 \phi^\dagger, \tag{15.2}$$

where ϕ is an $N = 1$ chiral multiplet and W_α is an $N = 1$ vector multiplet. The fact that $\phi^\dagger$ appears as the coefficient of the $\tilde{\theta}^2$ term is related to an additional constraint satisfied by ψ. This expression can be generalized to non-Abelian symmetries; the expression for the highest component of ψ is then somewhat more complicated; we won't need this here.

The theory possesses an $SU(2)$ R-symmetry under which θ and $\tilde{\theta}$ form a doublet. Under this symmetry, the scalar component of ϕ, and the gauge field, are singlets, while ψ and λ form a doublet.

I won't describe the hypermultiplets in detail, except to note that from the perspective of $N = 1$, they consist of two chiral multiplets. The two chiral multiplets transform as a doublet of the $SU(2)$. The superspace description of these multiplets is more complicated.

In the case of a non-Abelian theory, the vector field, ψ^a, is in the adjoint representation of the gauge group. For these fields, the Lagrangian has a very simple expression as an integral over half of the superspace:

$$\mathcal{L} = \int d^2\theta d^2\tilde{\theta}\ \psi^a \psi^a, \tag{15.3}$$

or, in terms of $N = 1$ components,

$$\mathcal{L} = \int d^2\theta\ W_\alpha^2 + \int d^4\theta \phi^\dagger e^V \phi. \tag{15.4}$$

The theory with vector fields alone has a classical moduli space, given by the values of the fields for which the scalar potential vanishes. Here this just means that the D fields vanish. Written as a matrix,

$$D = [\phi, \phi^\dagger], \tag{15.5}$$

which vanishes for diagonal ϕ, i.e. for

$$\phi = \frac{a}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{15.6}$$

For many physically interesting questions, one can focus on the effective theory for the light fields. In the present case, the light field is the vector multiplet, ψ. Roughly,

$$\psi \approx \psi^a \psi^a = a^2 + a\delta\psi^3 + \cdots . \tag{15.7}$$

What kind of effective action can we write for ψ? At the level of terms with up to four derivatives, the most general effective Lagrangian has the form:[1]

$$\mathcal{L} = \int d^2\theta d^2\tilde{\theta}\, f(\psi) + \int d^8\theta \mathcal{H}(\psi, \psi^\dagger). \tag{15.8}$$

Terms with covariant derivatives correspond to terms with more than four derivatives, when written in terms of ordinary component fields.

The first striking result we can read off from this Lagrangian, with no knowledge of $\mathcal{H}$ and f, is that there is no potential for ϕ, i.e. the moduli space is exact. This statement is true perturbatively and non-perturbatively.

One can next ask about the function f. This function determines the effective coupling in the low-energy theory, and is the object studied by Seiberg and Witten that we will discuss in Section 15.4.

15.2 A still simpler theory: $N = 4$ Yang–Mills

The $N = 4$ Yang–Mills theory is an interesting theory in its own right: it is finite and conformally invariant. It also plays an important role in our current understanding of non-perturbative aspects of string theory. The $N = 4$ Yang–Mills has 16 supercharges, and is even more tightly constrained than the $N = 2$ theories. First, we should describe the theory. In the language of $N = 2$ supersymmetry, it consists of one vector multiplet and one hypermultiplet. In terms of $N = 1$ superfields, it contains three chiral superfields, ϕ_i, and a vector multiplet. The Lagrangian is

$$\mathcal{L} = \int d^2\theta W_\alpha^2 + \int d^4\theta \phi_i^\dagger e^V \phi_i + \int d^2\theta \phi_i^a \phi_j^b \phi_k^c \epsilon_{ijk}\epsilon^{abc}. \tag{15.9}$$

In the above description, there is a manifest $SU(3) \times U(1)$ R-symmetry. Under this symmetry, the ϕ_is have $U(1)_R$ charge $2/3$, and form a triplet of the $SU(3)$. But the real symmetry is larger – it is $SU(4)$. Under this symmetry, the four Weyl fermions form a 4, while the 6 scalars transform in the 6. Thinking of these theories as the low-energy limits of toroidal compactifications of the heterotic string will later give us a heuristic understanding of this $SU(4)$: it reflects the $O(6)$ symmetry of the compactified 6 dimensions. In string theory, this symmetry is broken by

[1] This, and essentially all of the effective actions we will discuss, should be thought of as Wilsonian effective actions, obtained by integrating out heavy fields and high-momentum modes.

the compactification manifold; this reflects itself in higher-derivative, symmetry-breaking operators.

In the $N = 4$ theory, there is, again, no modification of the moduli space, perturbatively or non-perturbatively. This can be understood in a variety of ways. We can use the $N = 2$ description of the theory, defining the vector multiplet to contain the $N = 1$ vector and one (arbitrarily chosen) chiral multiplet. Then an identical argument to that given above insures that there is no superpotential for the chiral multiplet alone. The $SU(3)$ symmetry then insures that there is no superpotential for any of the chiral multiplets. Indeed, we can make an argument directly in the language of $N = 1$ supersymmetry. If we try to construct a superpotential for the low energy theory in the flat directions, it must be an $SU(3)$-invariant, holomorphic function of the ϕ_is. But there is no such object.

Similarly, it is easy to see that there are no corrections to the gauge couplings. For example, in the $N = 2$ language, we want to ask what sort of function, f, is allowed in

$$\mathcal{L} = \int d^2\theta d^2\tilde{\theta}\, f(\psi). \tag{15.10}$$

But the theory has a $U(1)$ R-invariance under which

$$\psi \to e^{2/3i\alpha}\psi \quad \theta \to e^{i\alpha\theta} \quad \tilde{\theta} \to e^{-i\alpha}\tilde{\theta}. \tag{15.11}$$

Already, then,

$$\int d^2\theta d^2\tilde{\theta}\,\psi\psi \tag{15.12}$$

is the unique structure which respects these symmetries. Now we can introduce a background dilaton field, τ. Classically, the theory is invariant under shifts in the real part of τ, $\tau \to \tau + \beta$. This insures that there are no perturbative corrections to the gauge couplings. More work is required to show that there are no non-perturbative corrections either.

One can also show that the quantity $\mathcal{H}$ in Eq. (15.8) is unique in this theory, again using the symmetries. The expression:

$$\mathcal{H} = c\ln(\psi)\ln(\psi^\dagger), \tag{15.13}$$

respects all of the symmetries. At first sight, it might appear to violate scale invariance; given that ψ is dimensionful, one would expect a scale, Λ, sitting in the logarithm. However, it is easy to see that if one integrates over the full superspace, any Λ-dependence disappears, since ψ is chiral. Similarly, if one considers the $U(1)$ R-transformation, the shift in the Lagrangian vanishes after the integration over superspace. To see that this expression is not renormalized, one merely needs to note that any non-trivial τ-dependence spoils these two properties. As a result, in

the case of $SU(2)$, the four derivative terms in the Lagrangian are not renormalized. Note that this argument is non-perturbative. It can be generalized to an even larger class of higher-dimension operators.

15.3 A deeper understanding of the BPS condition

In our study of monopoles, we saw that under certain circumstances, the complicated second-order non-linear differential equations reduced to first-order differential equations. The main condition is that the potential should vanish. We are now quite used to the idea that supersymmetric theories often have moduli, and have seen that this is an exact feature of $N = 4$ and many $N = 2$ theories. In the case of an $N = 2$ supersymmetric gauge theory, the potential is just that arising from the D-term, and one can construct a Prasad–Sommerfield solution. We will now see that the BPS condition is not simply magic, but is a consequence of the extended supersymmetry of the theory. The resulting mass formula, as a consequence, is *exact*; it is not simply a feature of the classical theory, but a property of the full quantum theory. This sort of BPS condition is relevant not only to the study of magnetic monopoles but to topological objects in various dimensions and contexts, particularly in string theory. Here we will give the flavor of the argument, without worrying carefully about factors of two. More details are worked out in the Exercises and the References.

First, we show that the electric and magnetic charges enter in the supersymmetry algebra of this theory as central charges. Thinking of this as an $N = 1$ theory, we have seen that the supercurrents take the form:

$$S^\mu_\alpha = \sigma^\mu_{\alpha\dot\beta}(\sigma^{\rho\sigma})^{\dot\beta\dot\gamma} F_{\rho\sigma}\lambda_{\dot\gamma} + \partial_\rho \phi^i \sigma^\rho_{\alpha\dot\beta}(\sigma^\mu)^{\dot\beta\gamma}\psi^i_\gamma + \text{ F-term pieces.} \qquad (15.14)$$

In this theory, however, there is an $SU(4)$ symmetry, and the supercurrents should transform as a 4. It is not hard to guess the others, even without writing down the transformation laws of the fields:

$$S^i_{\mu\alpha} = \sigma^\mu_{\alpha\dot\beta}(\sigma^{\rho\sigma})^{\dot\beta\dot\gamma} F_{\rho\sigma}\psi^i_{\dot\gamma} + \epsilon_{ijk}\partial_\rho \phi^j \sigma^\rho_{\alpha\dot\beta}(\sigma^\mu)^{\dot\beta\gamma}\psi^k_\gamma + \text{ F-term pieces.} \qquad (15.15)$$

We are interested in proving bounds on the mass. It is useful to define Hermitian combinations of the Qs, since we want to study positivity constraints. In this case, it is more convenient to write a four-component expression, using a Majorana (real) basis for the γ matrices. Taking an $N = 2$ subgroup, and carefully computing the commutators of the charges:

$$\{Q_{\alpha i}, Q_{\beta j}\} = \delta_{ij}\gamma^\mu_{\alpha\beta} P_\mu + \epsilon_{ij}(\delta_{\alpha\beta} U_k + (\gamma_5)_{\alpha\beta} V_k). \qquad (15.16)$$

Here

$$U_k = \int d^3x\, \partial_i \left(\phi^a_{\mathrm{re}\,k} E^a_i + \phi^a_{\mathrm{im}\,k} B^a_i\right)$$
$$V_k = \int d^3x\, \partial_i \left(\phi^a_{\mathrm{im}\,k} E^a_i + \phi^a_{\mathrm{re}\,k} B^a_i\right). \tag{15.17}$$

In the Higgs phase, the integrals are, by Gauss's theorem, electric and magnetic charges, multiplied by the Higgs expectation value.

From these relations, we can derive bounds on masses, using the fact that Q_α^2 is a positive operator. Taking the expectation of both sides, we have, for an electrically neutral system of mass M in its rest frame,

$$M \pm Q_{\mathrm{m}} v \geq 0. \tag{15.18}$$

This bound is saturated when Q annihilates the state. Examining the form of Q_α, this is just the BPS condition.

15.3.1 $N = 4$ *Yang–Mills theories and electric–magnetic duality*

This is a theory which, from the point of view of $N = 1$ supersymmetry, contains a gauge multiplet and three chiral multiplets in the adjoint representation. In addition to the interactions implied by the gauge symmetry, there is a superpotential:

$$W = \frac{1}{6} f_{abc} \epsilon_{ijk} \Phi^a_i \Phi^b_j \Phi^c_k. \tag{15.19}$$

We have normalized the kinetic terms for the fields Φ with a $1/g^2$ out front. So this interaction has strength related to the strength of the gauge interactions. This theory has a global $SU(4)$ symmetry. Under this symmetry, the 4 adjoint fermions transform as a 4; the scalars transform as a 6, and the gauge bosons are invariant. The theory has a large set of flat directions. If we simply take all of the Φ fields as matrices, to be diagonal the potential vanishes. As a result, this theory has monopoles of the BPS type.

This theory has a symmetry even larger than the Z_2 duality symmetry we contemplated when we examined Maxwell's equations; the full symmetry is $SL(2, Z)$. We might guess this, first, by remembering that the coupling constant is part of the holomorphic variable:

$$\tau = \frac{\theta}{2\pi} + \frac{4\pi i}{e^2}. \tag{15.20}$$

So in addition to our conjectured $e \to 1/e$ symmetry, there is a symmetry $\theta \to$

$\theta + 2\pi$. So in terms of τ, we have the two symmetry transformations,

$$\tau \to -\frac{1}{\tau} \quad \tau \to \tau + 1. \tag{15.21}$$

Together, these transformations generate the group $SL(2, Z)$:

$$\tau \to \frac{a\tau + b}{c\tau + d} \quad ad - bc = 1. \tag{15.22}$$

Now we can look at our BPS formula. To understand whether it respects the $SL(2, Z)$ symmetry we need to understand how this symmetry acts on the states. Writing

$$M = eQ_{\mathrm{e}}v + \frac{Q_{\mathrm{m}}v}{e} \tag{15.23}$$

with

$$Q_{\mathrm{e}} = n_{\mathrm{e}} - n_{\mathrm{m}}\frac{\theta}{2\pi} \quad Q_{\mathrm{m}} = 4\pi\frac{n_{\mathrm{m}}}{e} \tag{15.24}$$

the spectrum is invariant under the $SL(2, Z)$ transformation of τ, accompanied by:

$$\begin{pmatrix} n_{\mathrm{e}} \\ n_{\mathrm{m}} \end{pmatrix} \to \begin{pmatrix} d & -b \\ c & -d \end{pmatrix} \begin{pmatrix} n_{\mathrm{e}} \\ n_{\mathrm{m}} \end{pmatrix}. \tag{15.25}$$

Because it follows from the underlying supersymmetry, the mass formula is exact, so this duality of the spectrum of BPS objects is a non-perturbative statement about the theory.

15.4 Seiberg–Witten theory

We have seen that $N = 4$ theories are remarkably constrained, and this allowed us, for example, to explore an exact duality between electricity and magnetism. Still, these theories are not nearly as rich as field theories with $N \leq 1$ supersymmetry. The $N = 2$ theories are still quite constrained, but exhibit a much more interesting array of phenomena. They illustrate the power provided by supersymmetry over non-perturbative dynamics. They will also allow us to study phenomena associated with magnetic monopoles in a quite non-trivial way. In this section, we will provide a brief introduction to the subject known as *Seiberg–Witten theory*. This subject has applications not only in quantum field theory, but also for our understanding of string theory and, perhaps most dramatically, in mathematics.

It is convenient to describe the $N = 2$ theories in $N = 1$ language. The basic $N = 2$ multiplets are the vector multiplet and the tensor (or hyper) multiplet. From the point of view of $N = 1$ supersymmetry, the vector contains a vector multiplet

and a chiral multiplet. The tensor contains two chiral fields. We will focus mainly on theories with only vector multiplets, with gauge group $SU(2)$. In the $N = 1$ description, the fields are a vector multiplet, V, and a chiral multiplet, ϕ, both in the adjoint representation. The Lagrangian density is:

$$\mathcal{L} = \int d^4\theta \frac{1}{g^2}\phi^\dagger e^V \phi - \frac{i}{16\pi}\int d^2\theta \tau W^{a\alpha} W^a_\alpha + \text{h.c.} \tag{15.26}$$

Here

$$\tau = \frac{\theta}{2\pi} + i4\pi g^{-2}. \tag{15.27}$$

The $1/g^2$ in front of the chiral field kinetic term is somewhat unconventional, but it makes the $N = 2$ supersymmetry more obvious. As we indicated earlier, one way to understand the $N = 2$ supersymmetry is to note that the Lagrangian we have written has a global $SU(2)$ symmetry. Under the symmetry, the scalar fields, ϕ^a, and the gauge fields A^a_μ are singlets, while the gauginos, λ^a, and the fermionic components of ϕ, ψ^a, transform as a doublet. Acting on the conventional $N = 1$ generators, the $SU(2)$ produces four new generators. So we have generators Q^A_α, with $A = 1, 2$.

As it stands, the model has flat directions, with

$$\phi = \frac{a}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{15.28}$$

In these directions, the spectrum consists of two massive and one massless gauge bosons, a massive complex scalar, degenerate with the gauge bosons, and a massive Dirac fermion, as well as a massless vector and a massless chiral multiplet. The masses of all of these particles are

$$M_W = \sqrt{2}a. \tag{15.29}$$

This is precisely the right number of states to fill an $N = 2$ multiplet. Actually, this multiplet is a BPS multiplet. It is annihilated by half of the supersymmetry generators. The classical theory possesses, in addition to the global $SU(2)$ symmetry, an anomalous $U(1)$ symmetry:

$$\phi \to e^{i\alpha\phi} \quad \psi \to e^{i\alpha}\psi. \tag{15.30}$$

Under this symmetry,

$$\theta \to \theta - 4\alpha \tag{15.31}$$

or

$$\tau \to \tau - 2\pi\alpha. \tag{15.32}$$

Because physics is periodic in θ with period 2π, $\alpha = \pi/2$ is a symmetry, i.e. the theory has a Z_4 symmetry:

$$\phi \to e^{\frac{i\pi}{2}}\phi. \tag{15.33}$$

Note that ϕ is not gauge-invariant. A suitable gauge-invariant variable for the analysis of this theory is:

$$u = \langle \mathrm{Tr}\, \phi^3 \rangle. \tag{15.34}$$

Under the discrete symmetry, $u \to -u$; at weak coupling,

$$u \approx a^2. \tag{15.35}$$

The spectrum of this theory includes magnetic monopoles, in general with electric charges. At the classical level, the monopole solutions in this theory are precisely those of Prasad and Sommerfield, with mass

$$M_{\mathrm{M}} = 4\pi\sqrt{2}\frac{a}{g^2}. \tag{15.36}$$

As in the $N = 4$ theory, there is a BPS formula for the masses:

$$m = \sqrt{2}|a\,Q_{\mathrm{e}} + a_{\mathrm{D}}\,Q_{\mathrm{M}}|. \tag{15.37}$$

At tree level,

$$a_{\mathrm{D}} = \frac{4\pi}{g^2} ia = \tau a, \tag{15.38}$$

where the last equation holds if $\theta = 0$. The i in this formula is not immediately obvious. To see that it must be present, consider the case of dyonic excitations of monopoles. These should have energy of order the charge, with no factors of $1/g^2$. This is insured by the relative phase between a and a_{D}. These formulas will receive corrections in perturbation theory and beyond; our goal is to understand the form of these corrections and their (dramatic) physical implications.

Equation (15.38) is not even meaningful as it stands; τ is a function of scale. Instead, Seiberg and Witten suggested that

$$\tau = \frac{da_{\mathrm{D}}}{da}. \tag{15.39}$$

They also proposed the existence of a duality symmetry, under which

$$a_{\mathrm{D}} \leftrightarrow a \quad \tau \to -\frac{1}{\tau}. \tag{15.40}$$

To formulate our questions more precisely, and to investigate this proposal, it is helpful, as always, to consider a low-energy effective action. This action should respect the $N = 2$ supersymmetry; in $N = 1$ language, this means that the Lagrangian

should take the form:

$$\mathcal{L} = \int d^4\theta K(a, \bar{a}) - \frac{i}{16\pi^2} \int d^2\theta \tau(a) W^\alpha W_\alpha. \tag{15.41}$$

The $N = 2$ supersymmetry implies a relation between K and τ; without it, these would be independent quantities. Both quantities can be obtained from a holomorphic function called the prepotential, $\mathcal{F}(a)$:

$$\tau = \frac{d^2\mathcal{F}}{da^2} \quad K = \frac{1}{4\pi} \frac{d\mathcal{F}}{da} a^*. \tag{15.42}$$

From

$$\tau = \frac{da_{\mathrm{D}}}{da} = \frac{d}{da}\left(\frac{d\mathcal{F}}{da}\right) \tag{15.43}$$

we have

$$\frac{d\mathcal{F}}{da} = i a_{\mathrm{D}}, \tag{15.44}$$

so

$$K = \frac{1}{4\pi} \operatorname{Im} a_{\mathrm{D}} a^*. \tag{15.45}$$

Our goal will be to obtain a non-perturbative description of $\mathcal{F}$. At weak coupling, the beta function of this theory is obtained from $b_0 = 3N - N = 2N = 4$, so

$$\tau = \frac{i}{\pi} \ln(u/\Lambda^2). \tag{15.46}$$

As a check on this formula, note that, under $u \to e^{2i\alpha} u$, $\theta \to \theta - 4\alpha$, so

$$\tau = \frac{\theta}{2\pi} + 4\pi i g^{-2} \to \tau - \frac{2i\alpha}{\pi} \tag{15.47}$$

and this is precisely the behavior of the formula (Eq. (15.46)).

This is similar to phenomena we have seen in $N = 1$ theories. But when we consider the monopoles of the theory, the situation becomes more interesting. First, note that, using the leading order result for τ,

$$a_{\mathrm{D}} = \frac{2i}{\pi}\left(a \ln\left(\frac{a}{\Lambda}\right) - a\right). \tag{15.48}$$

So, under the transformation of u, $u \to e^{i\alpha u}$,

$$a_{\mathrm{D}} \to e^{i\alpha/4}\left(a_{\mathrm{D}} - \frac{\alpha}{2\pi} a\right). \tag{15.49}$$

Our BPS mass formula transforms to:

$$m \to \sqrt{2}\left|a\left(Q_{\mathrm{e}} - \frac{4\alpha}{2\pi} Q_{\mathrm{m}}\right) + a_{\mathrm{D}} Q_{\mathrm{M}}\right|. \tag{15.50}$$

This is the Witten effect, which we have discussed earlier: in the presence of θ, a magnetic monopole acquires an electric charge. More generally, the spectrum of dyons is altered.

Consider now what happens when we do a full 2π change of θ ($u \to -u$). This should be a symmetry. This is true in this case, but in a subtle way: the spectrum of the dyonic excitations of the theory is unchanged, but the charges of the dyons have shifted by one fundamental unit. This, in turn, is related to the branched structure of τ.

At the non-perturbative level, the structure is even richer. We might expect

$$\tau(u) = \frac{i}{\pi} \ln(u/\Lambda^2) + \alpha e^{-\frac{8\pi^2}{g^2}} + \beta e^{-\frac{8\pi^2}{g^2}} + \cdots . \tag{15.51}$$

Note that, interpreting $\exp(-8\pi^2/g^2)$ as $\exp(2\pi i \tau)$, each term in this series has the correct periodicity in θ. Moreover,

$$\exp(2\pi i \tau) = \frac{\Lambda^2}{u^2}. \tag{15.52}$$

These corrections have precisely the correct structure to be instanton corrections, and these instanton corrections have been computed. But we can, following Seiberg and Witten, be bolder and consider what happens when g becomes large. Naively, we might expect that some monopoles become light. Associated with this, τ may have a singularity at some point, $u_0 = \gamma \Lambda^2$, where Λ is the renormalization group invariant mass of the theory. In light of the Z_2 symmetry, there must also be a singularity at $-u_0$. Such a singularity arises because a particle is becoming massless. If we think of $\tau_{\rm D}$ as the dual of τ, then there is an electrically charged light field of unit charge; more precisely, there must be two particles of opposite charge, in order that they can gain mass. So $\tau_{\rm D}$ has the structure:

$$\tau_{\rm D} = -\frac{2i}{2\pi} \ln(m_{\rm M}). \tag{15.53}$$

Assuming that $a_{\rm D}$ has a simple zero,

$$a_{\rm D} \approx b(u - u_0); \quad m_{\rm M} = \sqrt{2} a_{\rm D}, \tag{15.54}$$

then

$$\tau_{\rm D} = -\frac{i}{\pi} \ln(u - u_0) = -\frac{1}{\tau(u)}. \tag{15.55}$$

Starting with the relation

$$\frac{da}{da_{\rm D}} = -\tau_{\rm D} = -\frac{i}{\pi} \ln(a_{\rm D}) \tag{15.56}$$

we have

$$a = \frac{i}{\pi}(a_{\rm D} \ln a_{\rm D} - a_{\rm D}). \tag{15.57}$$

Similarly, we can consider the behavior at the point $-u_0$. This is the mirror of the previous case, but we must be careful about the relation of a, $a_{\rm D}$. These are connected by the symmetry transformation:

$$\tilde{a} = ia \quad \tilde{a}_{\rm D} = i(a_{\rm D} - a). \tag{15.58}$$

Now:

$$\tau_{\rm D} = -\frac{1}{\tau(u)} = -\frac{i}{\pi} \ln(u + u_0) \tag{15.59}$$

and

$$\tilde{a} = \frac{1}{\pi}(\tilde{a}_{\rm D} \ln \tilde{a}_{\rm D} - \tilde{a}_{\rm D}). \tag{15.60}$$

Going around the singularities, at u_0,

$$a \to a - 2a_{\rm D}; \;\; a_{\rm D} \to a_{\rm D}, \tag{15.61}$$

while at $-u_0$,

$$a \to 3a - 2a_{\rm D}; a_{\rm D} \to 2a - a_{\rm D}. \tag{15.62}$$

This should be compared with the effect of going around 2π at large u: $a \to -a$; $a_{\rm D} \to -(a_{\rm D} - a)$. Assuming that these are the only singularities, we can, from this information, reconstruct τ. We won't give the full solution of Seiberg and Witten here, but the basic idea is to note that $\tau(u)$ is the modular parameter of a two-dimensional torus and to reconstruct the torus.

This analysis has allowed us to study the theory deep in the non-perturbative region. Seiberg and Witten uncovered a non-trivial duality, a limit in which monopoles become massless, and provided insight into confinement. These sorts of ideas have been extended to other theories, to theories in higher dimensions, and have provided insight into many phenomena in string theory, quantum gravity and pure mathematics.

Suggested reading

The lectures by Lykken (1996) provide a brief introduction to aspects of $N > 1$ supersymmetry. Olive and Witten (1978) first clarified the connection of the BPS condition and extended supersymmetry, in a short and quite readable paper. Harvey (1996) provides a more extensive introduction to monopoles and the BPS condition.

The original papers of Seiberg and Witten (1994) are quite clear; Peskin's lectures, which we have borrowed from extensively here, provide a brief and very clear introduction to the subject.

Exercises

(1) Check the supersymmetry commutators in extended supersymmetry (Eq. (15.16)).
(2) Rewrite the supersymmetry commutators in a real basis for the Dirac matrices. Using this, verify the BPS inequality.
(3) Check that the monopole/dyon spectrum in Eq. (15.23) is invariant under $SL(2, Z)$ transformations.

16

More supersymmetric dynamics

While motivated in part by the hopes of building phenomenologically successful models of particle physics, we have uncovered in our study of supersymmetric theories a rich trove of field theory phenomena. Supersymmetry provides powerful constraints on dynamics. In this chapter, we will discover more remarkable features of supersymmetric field theories. We will first study classes of (super)conformally invariant field theories. Then we will turn to the dynamics of supersymmetric QCD with $N_f \geq N_c$, where we will encounter new, and rather unfamiliar, types of behavior.

16.1 Conformally invariant field theories

In quantum field theory, theories which are classically scale-invariant typically are not scale invariant at the quantum level. QCD is a familiar example. In the absence of quark masses, we believe the theory confines and has a mass gap. The CP^N models are an example where we were able to show systematically how a mass gap can arise in a scale-invariant theory. The breaking of scale invariance in all of these cases is associated with the need to impose a cutoff on the high-energy behavior of the theory. In a more Wilsonian language, one needs to specify a scale where the theory is defined, and this requirement breaks the scale invariance.

There is, however, a subset of field theories which are scale invariant. We have seen this in the case of $N = 4$ supersymmetric field theories in four dimensions. In this section, we will see that this phenomenon can occur in $N = 1$ theories, and explore some of its consequences. In the next section we will discuss a set of dualities among $N = 1$ supersymmetric field theories, in which conformal invariance plays a crucial role.

In order that a theory exhibit conformal invariance, it is necessary that its beta (β-)function vanish. At first sight, it would seem difficult to use perturbation theory to find such theories. For example, one might try to choose the number of flavors

and colors so that the one-loop beta function vanishes. But then the two-loop beta function will generally not vanish. One could try to balance the first term against the second, but this would generally require $g^4 \sim g^2$, and there would not be a good perturbation expansion. Banks and Zaks pointed out that one can find such theories by adopting a different strategy. By taking the number of flavors and colors large, one can arrange that the coefficient of the one-loop beta function almost vanishes, and choose the coupling so that it cancels the two-loop beta function. In this situation, one can arrange a cancellation perturbatively, order by order. The small parameter is $1/N$, where N is the number of colors.

We can illustrate this idea in the framework of supersymmetric theories with N colors and $N_{\rm f}$ flavors. The beta function, through two loops, is given by

$$\beta(g) = -\frac{g^3}{16\pi^2} b_0 - \frac{g^5}{(16\pi^2)^2} b_1 \tag{16.1}$$

where

$$b_0 = 3N - N_{\rm f} \quad b_1 = 6N^2 - 2NN_F + 4N_F \frac{(N^2-1)}{2N}. \tag{16.2}$$

In the limit of very large N, $N_{\rm f}$, we write: $N_{\rm f} = 3N - \epsilon$, where ϵ is an integer of order one. Then, to leading order in $1/N$, the beta function vanishes for a particular coupling, g_0:

$$\frac{g_0^2}{16\pi^2} = \frac{\epsilon}{6N^2}. \tag{16.3}$$

Perturbative diagrams behave as $(g^2N)^n$, and g^2N is small. So at each order, one can make small adjustments in g^2 so as to make the beta function vanish.

A theory in which the beta function vanishes is genuinely conformally invariant. We will not give a detailed discussion of the conformal group here; the exercises at the end of this chapter guide the reader through some of the features of the conformal group; good reviews are described in the suggested reading. Here we will just mention a few general features, and then perform some computations at our Banks–Zaks fixed point theories to verify these.

Without supersymmetry, the generators of the conformal group include the Lorentz generators and the translations:

$$M_{\mu\nu} = -i(x_\mu \partial_\nu - x_\nu \partial_\mu) \quad P_\mu = -i\partial_\mu \tag{16.4}$$

and the generators of "special conformal transformations" and dilatations:

$$K_\mu = -i(x^2 \partial_\mu - 2x_\mu x_\alpha \partial^\alpha) \quad D = i x_\alpha \partial^\alpha. \tag{16.5}$$

In the presence of supersymmetry, the group is enlarged. In addition to the bosonic generators above and the supersymmetry generators, there are a group

of *superconformal* generators, of the form

$$S_\alpha = X_\mu \sigma^\mu_{\alpha\dot{\alpha}} Q^{\dot{\alpha}}. \tag{16.6}$$

We encountered these in our analysis of the zero modes of the Yang–Mills instanton. The superconformal algebra also includes an R-symmetry current.

Conformal invariance implies the vanishing of T^μ_μ. In the superconformal case, the superconformal generators and the divergence of the R current also vanish. One can prove a relation between the dimension and the R charge:

$$d \geq \frac{3}{2}|R|. \tag{16.7}$$

States for which the inequality is satisfied are known as *chiral primaries*. An interesting case is provided by the fixed point theories we have introduced above. For these, the charge of the chiral fields, Q and $\bar{Q}$, under the *non-anomalous* symmetry is

$$R_{Q,\bar{Q}} = \frac{N_{\rm f} - N}{N_{\rm f}}. \tag{16.8}$$

Assuming that these fields are chiral primaries, it follows that their dimension satisfies:

$$d - 1 = -\frac{3N - N_{\rm f}}{2N_{\rm f}} = -\frac{\epsilon}{6N}. \tag{16.9}$$

At weak coupling, on the other hand, the anomalous dimensions of these fields are known:

$$\gamma = -\frac{g^2}{16\pi^2} N = -\frac{\epsilon}{6N}. \tag{16.10}$$

In this chapter, we will see that supersymmetric QCD, for a range of $N_{\rm f}$ and N, exhibits conformal fixed points for which the coupling is not small.

16.2 More supersymmetric QCD

We have studied the dynamics of supersymmetric QCD with $N_{\rm f} < N$, and observed a range of phenomena: non-perturbative effects which lift the degeneracy among different vacua and non-perturbative supersymmetry breaking. In the cases $N_{\rm f} \geq N_{\rm c}$, there are exact moduli, even non-perturbatively. For phenomenology, such theories are probably of no relevance, but Seiberg realized that from a theoretical point of view, these theories are a bonanza. The existence of moduli implies a great deal of control over the dynamics. One can understand much about the strongly coupled regimes of these theories, allowing insights into non-perturbative dynamics unavailable in theories without supersymmetry. We will be able to answer questions

such as: are there unbroken global symmetries in some region of the moduli space? In regions of strong coupling, are there massless composite particles?

16.3 $N_f = N_c$

The case $N_f = N_c$ already raises issues beyond those of $N_f < N_c$. First, we have seen that there is no invariant superpotential one can write. As a result, there is an exact moduli space, perturbatively and non-perturbatively. But there is still an interesting quantum modification of the theory, first discussed by Seiberg.

Consider, first, the classical moduli space. Now, in addition to the vacua with $Q = \bar{Q}$ (up to flavor transformations) which we found previously, we can also have

$$Q = vI \quad \bar{Q} = 0 \text{ or } Q \leftrightarrow \bar{Q}. \tag{16.11}$$

This is referred to as the "baryonic branch," since now the operator:

$$B = \epsilon_{i_1 \ldots i_N} \epsilon^{j_1 \ldots j_N} Q^{i_1}_{j_1} \ldots Q^{i_N}_{j_N} \tag{16.12}$$

is non-vanishing (or the corresponding "anti-baryon").

Classically, these two classes of possibilities can be summarized in the condition:

$$\text{Det}(\bar{Q}Q) = \bar{B}B. \tag{16.13}$$

Now this condition is subject to quantum modifications. Both sides are completely neutral under the various flavor symmetries; in principle any function of $B\bar{B}$ or the determinant would be permitted as a modification. But we can use anomalous symmetries (with the anomalies canceled by shifts in S) to constrain any possible corrections. Consider, in particular, possible instanton corrections. These are proportional to

$$v^{2N} e^{\frac{-8\pi^2}{g^2(v)}} \sim \Lambda^{2N} \tag{16.14}$$

and transform just like the left-hand side under the anomalous R-symmetry for which:

$$Q \to e^{i\alpha} Q. \tag{16.15}$$

So at the quantum level, the moduli space satisfies the condition:

$$\text{Det}(\bar{Q}Q) - \bar{B}B = c\Lambda^{2N}. \tag{16.16}$$

This is of just the right form to be generated by a 1-instanton correction. We will not do the calculation here which shows that the right-hand side *is* generated. But we can outline the main features. There are now 2 superconformal zero modes, 2 supersymmetry zero modes, $4N - 4$ zero modes associated with the

gluinos in the $(2, N-2)$ representation of the $SU(2) \times SU(N-2)$ subgroup of $SU(N)$ distinguished by the instanton, and $2N$ matter zero modes. We want to compute the expectation value of an operator involving N scalars. To obtain a non-vanishing result, it is necessary to replace some of these fields with their classical values. Others must be contracted with Yukawa terms. The scalar field propagators in the instanton background are known, and the full calculation reasonably straightforward. Because the classical condition which defines the moduli space is modified, the moduli space of the $N_f = N_c$ theory is referred to as the "quantum moduli space." This phenomenon appears for other choices of gauge group as well.

16.3.1 Supersymmetry breaking in quantum moduli spaces

We have mentioned that in the 3−2 model, in the limit that the $SU(2)$ gauge group is the strong group, supersymmetry breaking can be understood as resulting from an expectation value for QL. The QL vev is non-zero since $N = N_f = 2$. A larger class of models in which a quantum moduli space is responsible for dynamical supersymmetry is due to Intriligator and Thomas.

Consider a model with gauge group $SU(2)$ and four doublets, Q_I, $I = 1 \ldots 4$ (two "flavors"). Classically, this model has a moduli space labelled by the expectation values of the fields $M_{IJ} = Q_I Q_J$. These satisfy $\mathrm{Pf}\langle M_{IJ}\rangle = 0$,[1] but, as have have just seen, the quantum moduli space is different, and satisfies:

$$\mathrm{Pf}\langle M_{IJ}\rangle = \Lambda^4. \tag{16.17}$$

Now add a set of singlets to the model, S_{IJ}, with superpotential couplings

$$W = \lambda_{IJ} S_{IJ} Q_I Q_J. \tag{16.18}$$

Unbroken supersymmetry now requires

$$\frac{\partial W}{\partial S_{IJ}} = Q_I Q_J = 0. \tag{16.19}$$

However, this is incompatible with the quantum constraint. So supersymmetry is broken.

On the other hand, the model, classically, has flat directions in which $S_{IJ} = s_{IJ}$, and all of the other fields vanish. So one might worry that there is runaway behavior in these directions, similar to that we saw in supersymmetric QCD. However, for large s, it turns out that the energy grows at infinity. This can be established as

[1] In this expression, Pf denotes Pfaffian. The Pfaffian is defined for $2N \times 2N$ antisymmetric matrices; it is essentially the square root of the determinant of the matrix.

follows. Suppose all of the components of S are large, $S \sim s \gg \Lambda_2$. In this limit, the low-energy theory is a pure $SU(2)$ gauge theory. In this theory, gluinos condense,

$$\langle \lambda\lambda \rangle = \Lambda_{\rm LE}^3 = \lambda s \Lambda_2^2. \tag{16.20}$$

Here, $\Lambda_{\rm LE}$ is the Λ parameter of the low-energy theory.

At this level, then, the superpotential of the model behaves as

$$W_{\rm eff} \sim \lambda S \Lambda_2^2, \tag{16.21}$$

and the potential is a constant,

$$V = |\Lambda_2|^4 |\lambda|^2. \tag{16.22}$$

The natural scale for the coupling, λ, which appears here is $\lambda(s)$. This is the correct answer in this case, and implies that for large s the potential grows, since λ is not asymptotically free. So the potential has a minimum, in a region of small s.

16.3.2 $N_{\rm f} = N_{\rm c} + 1$

For $N_{\rm f} > N_{\rm c}$, the classical moduli space is exact. But Seiberg has, again, pointed out a rich set of phenomena and given a classification of the different theories. As in the case of $N_{\rm f} < N_{\rm c}$, different phenomena occur for different values of $N_{\rm f}$.

First, we need to introduce a new tool: the 't Hooft anomaly matching conditions. 't Hooft was motivated by the following question. When one looks at the repetitive structure of the quark and lepton generations, it is natural to wonder if the quarks and leptons themselves are bound states of some simpler constituents. 't Hooft pointed out that if this idea is correct, the masses of the quarks and leptons are far smaller than the scale of the underlying interactions; even at that time, it was known that if these particles have any structure, it is on scales shorter than 100 GeV^{-1}. 't Hooft argued that this could only be understood if the underlying interactions left an unbroken chiral symmetry.

One could go on and simply postulate that the symmetry is unbroken, but 't Hooft realized that there are strong – and simple – constraints on such a possibility. Assuming that the underlying interactions are some strongly interacting non-Abelian gauge theory, 't Hooft imagined gauging the global symmetries of the theory. In general, the resulting theory would be anomalous, but one could always cancel the anomalies by adding some "spectator" fields, fields transforming under the gauged flavor symmetries but not the underlying strong interactions. Below the confinement scale of the strong interactions, the flavor symmetries might be spontaneously broken, giving rise to Goldstone bosons, or there might be massless fermions. In either case, the low-energy theory must be anomaly-free, so the anomalies of either the Goldstone bosons or the massless fermions must be the same as in the

original theory. 't Hooft added another condition, which he called the "decoupling" condition: he asked what happened if one added mass terms for some of the constituent fermions. He went on to show that these conditions are quite powerful, and that it is difficult to obtain a theory with unbroken chiral symmetries.

As we will see, Seiberg conjectured various patterns of unbroken symmetries for SUSY QCD. For these the 't Hooft anomaly conditions provide a strong self-consistency check. In the case $N_f = N_c$, there is no point in the moduli space where the chiral symmetries are all unbroken. So let's move on to the case $N_f = N_c + 1$. The global symmetry of the model is:

$$SU(N_f)_L \times SU(N_f)_R \times U(1)_B \times U(1)_R \tag{16.23}$$

where under $U(1)_R$, the quarks and antiquarks transform as:

$$Q_f, \bar{Q}_f \to e^{\frac{i\alpha}{N+1}} Q_f, \bar{Q}_f. \tag{16.24}$$

In this theory, there two various sorts of gauge invariant objects: the mesons, $M_{\bar{f}f} = \bar{Q}_{\bar{f}} Q$ and the baryons, $B_f = \epsilon_f^{\alpha_1 \cdots \alpha_N} \epsilon_{i_1 \cdots i_n} Q^{i_1}_{\alpha_1} Q^{i_2}_{\alpha_2} \cdots Q^{i_N}_{\alpha_N}$. From these, we can build a superpotential invariant under all of the symmetries:

$$W = (\det M - B_{\bar{f}} M_{\bar{f}f} B_f) \frac{1}{\Lambda^{b_0}}. \tag{16.25}$$

As in all of our earlier cases, the power of Λ is determined by dimensional arguments, but can also be verified by demanding holomorphy in the gauge coupling.

This superpotential has several interesting features. First, it has flat directions, as we would expect, corresponding to the flat directions of the underlying theory. But also, for the first time, there is a vacuum at the point where all of the fields vanish, $B = \bar{B} = M = 0$. At this point, all of the symmetries are unbroken. The 't Hooft anomaly conditions provide an important consistency check on this whole picture. There are several anomalies to check ($SU(N_f)^3_L$, $SU(N_f)^3_R$, $SU(N_f)^2_L U(1)_R$, $\mathrm{Tr} U(1)_R$, $U(1)^2_B U(1)_R$, $U(1)^3_R$, etc.). The cancellations are quite non-trivial. In the exercises, the reader will have the opportunity to check these.

Another test comes from considering decoupling. If we add a mass for one set of fields, the theory should reduce to the $N_f = N$ case. As in examples with smaller numbers of fields, we take advantage of holomorphy, writing expressions for small values of the mass and continuing to large values. So we add to the superpotential a term:

$$m \bar{Q}_{N+1} Q_{N+1} = m M_{N+1,N+1}. \tag{16.26}$$

We want to integrate out the massive fields. Because of the global symmetry, it is consistent to set $M_{f,N+1}$ to zero, where $f \leq N$. Similarly, it is consistent to set

$B_f = 0$, $f \leq N$. So we take the M and B fields to have the form:

$$M = \begin{pmatrix} M & 0 \\ 0 & m \end{pmatrix} \quad B_f = \begin{pmatrix} 0 \\ \cdots \\ B \end{pmatrix} \quad \bar{B}_f = \begin{pmatrix} 0 \\ \cdots \\ \bar{B} \end{pmatrix}. \tag{16.27}$$

Consider the equation $\partial W/\partial m = 0$. This yields:

$$(\det M - \bar{B}B) = m\Lambda^{b_0} \tag{16.28}$$

or

$$(\det \mathrm{M} - \bar{B}B) = m\Lambda^{b_0} = \Lambda_{N_\mathrm{f}}^{2N}. \tag{16.29}$$

In the last line, we have used the relation between the Λ parameter of the theory with N_f quarks and that with $N_\mathrm{f} + 1$ flavors. This is precisely the expression for the quantum modified moduli space of the N-flavor theory. Decoupling works perfectly here.

16.4 $N_\mathrm{f} > N + 1$

The case $N_\mathrm{f} > N + 1$ poses new challenges. We might try to generalize our analysis of the previous section. Take, for example, $N_\mathrm{f} = N + 2$. Then the baryons are in the second-rank antisymmetric tensor representations of the $SU(N_\mathrm{f})$ gauge groups, B_{fg}, $\bar{B}_{\bar{f}\bar{g}}$. If we write a term in the superpotential

$$W \sim B_{fg}\bar{B}_{\bar{k}\bar{l}}M^{f\bar{k}}M^{g\bar{l}} \tag{16.30}$$

this does not respect the non-anomalous R symmetry.

Seiberg suggested a different equivalence. The baryons, in general, have $\tilde{N} = N_\mathrm{f} - N$ indices. So baryons in the same representation of the flavor group can be constructed in a theory with gauge group $SU(\tilde{N})$ and quarks $q_f, \bar{q}_f$ in the fundamental representation. Seiberg postulated that, in the infrared, this theory is dual to the original theory. This is not quite enough. One needs to add a set of gauge-singlet meson fields, $M_{\bar{f}f}$, with superpotential:

$$W = q^{\bar{f}} M_{\bar{f}f} q^f. \tag{16.31}$$

To check this picture, we can first check that the symmetries match. There is an obvious $SU(N_\mathrm{f})_\mathrm{L} \times SU(N_\mathrm{f})_f \times U(1)_B$. There is also a non-anomalous $U(1)_R$ symmetry. It is important that the dual theory not be asymptotically free, i.e. that it is weakly coupled in the infrared. This is the case $N > 3N_\mathrm{f}/2$. So again, this duality can only apply for a range of N_f, N.

There are a number of checks on the consistency of this picture. Holomorphic decoupling is again one of the most persuasive. Take the case $N_f = N + 2$, so that the dual gauge group is $SU(2)$. In this case, working in the flat directions of the $SU(2)$ theory, one can do an instanton computation. One finds a contribution to the superpotential:

$$W_{\rm inst} = \det M. \tag{16.32}$$

This is consistent with all of the symmetries; it is not difficult to see that one can close up all of the fermion zero modes with elements of M and q. So one has a superpotential:

$$\int d^2\theta (q M \bar{q} - \det M). \tag{16.33}$$

16.5 $N_f \geq 3/2N$

We have noted that Seiberg's duality can't persist beyond $N_f = 3/2N$. Seiberg also made a proposal for the behavior of the theory in this regime: for $3/2N \leq N_f \leq 3N$, the theories are conformally invariant. Our Banks–Zaks fixed point lies in one corner of this range. As a further piece of evidence, consider the dimension of the operator $\bar{Q}Q$. Under the non-anomalous R-symmetry,

$$Q \to e^{i\alpha \frac{N_f - N}{2N_f}} Q. \tag{16.34}$$

If the theory is superconformal, the dimension of this chiral operator satisfies $d = 3/2R$. As explained in Appendix D, the exact beta function of the theory is:

$$\beta = -\frac{g^3}{16\pi^2} \frac{3N - N_F + N_f \gamma(g^2)}{1 - N(g^2/8\pi^2)}. \tag{16.35}$$

By assumption, this is zero, so

$$\gamma = -\frac{3N - N_f}{N_f}. \tag{16.36}$$

The dimension of $\bar{Q}Q$ is $2 + \gamma$, which is precisely $3/2R$.

We will not pursue this subject further, but there is further evidence one can provide for all of these dualities. They can also be extended to other gauge groups.

Suggested reading

The original papers of Seiberg (1994a,b, 1995a,b; see also Seiberg and Witten 1994) are quite accessible and essential reading on these topics, as is the review by Intriligator and Seiberg (1996). Good introductions are provided by the lecture

notes of Peskin (1997) and Terning (2003). The use of quantum moduli spaces to break supersymmetry was introduced in Intriligator and Thomas (1996).

Exercises

(1) Discuss the renormalization of the composite operator $\bar{Q}Q$, and verify that the relation $d = 3/2R$ is again satisfied.
(2) Check the anomaly cancellation for the case $N_f = N + 1$. You may want to use an algebraic manipulation program, like MAPLE or Mathematica, to expedite the algebra.

17

An introduction to general relativity

Even as the evidence for the Standard Model became stronger and stronger in the 1970s and beyond, so the evidence for general relativity grew in the latter half of the twentieth century. Any discussion of the Standard Model and physics beyond must confront Einstein's theory at two levels. First, general relativity and the Standard Model are very successful at describing the history of the universe and its present behavior on large scales. General relativity gives rise to the big bang theory of cosmology, which, coupled with our understanding of atomic and nuclear physics, explains – indeed predicted – features of the observed universe. But there are features of the observed universe which cannot be accounted for within the Standard Model and general relativity. These include the dark matter and the dark energy, the origin of the asymmetry between matter and antimatter, the origin of the seeds of cosmic structure (inflation), and more. Apart from these observational difficulties, there are also serious questions of principle. We cannot simply add Einstein's theory onto the Standard Model. The resulting structure is not renormalizable, and cannot represent in any sense a complete theory. In this book we will encounter both of these aspects of Einstein's theory. Within extensions of the Standard Model, in the next few chapters, we will attempt to explain some of the features of the observed universe. The second, more theoretical, level, is addressed in the third part of this book. String theory, our most promising framework for a comprehensive theory of all interactions, encompasses general relativity in an essential way; some would even argue that what we mean by string theory is the quantum theory of general relativity.

The purpose of this chapter is to introduce some of the concepts and formulas that are essential to the applications of general relativity in this text. No previous knowledge of general relativity is assumed. We will approach the subject from the perspective of field theory, focussing on the dynamical degrees of freedom and the equations of motion. We will not give as much attention to the beautiful – and conceptually critical – geometric aspects of the subject, though we will return to

some of these in the chapters on string theory. Those interested in more serious study of general relativity will eventually want to study some of the excellent texts listed in the suggested reading at the end of the chapter.

Einstein put forward his principle of relativity in 1905. At that time, one might quip that half the known laws, those of electricity and magnetism, already satisfied this principle, with no modification. The other half, Newton's laws, did not. In considering how one might reconcile gravitation and special relativity, Einstein was guided by the observed equality of gravitational and inertial mass. Inertia has to do with how objects move in space-time in response to forces. Operationally, the way we define space-time – our measurements of length, time, energy and momentum – depends crucially on this notion. The fact that gravity couples to precisely this mass suggests that gravity has a deep connection to the nature of space-time. Considering this equivalence, Einstein noted that an observer in a freely falling elevator (in a uniform gravitational field) would write down the same laws of nature as an observer in an inertial frame without gravity. Consider, for example, an elevator full of particles interacting through a potential $V(\vec{x}_i - \vec{x}_j)$. In the inertial frame,

$$m\frac{d^2\vec{x}_i}{dt^2} = m\vec{g} - \vec{\nabla}_i V(\vec{x}_i - \vec{x}_j). \tag{17.1}$$

The coordinates of the accelerated observer are related to those of the inertial observer by

$$\vec{x}_i = \vec{x}'_i + \frac{1}{2}\vec{g}t^2 \tag{17.2}$$

so, plugging in the equations of motion:

$$m\frac{d^2\vec{x}'_i}{dt^2} = -\vec{\nabla}_i V(\vec{x}'_i - \vec{x}'_j). \tag{17.3}$$

Einstein abstracted from this thought experiment a strong version of the *equivalence principle*: the equations of motion should have the same form in any frame, inertial or not. In other words, it should be possible to write the laws so that in two coordinate systems, x^μ and $x'^\mu(x)$, they take the same form. This is a strong requirement. We will see that it is similar to gauge invariance, where the requirement that the laws take the same form after gauge transformations determines the dynamics.

17.1 Tensors in general relativity

To implement the equivalence principle, we begin by thinking about the invariant element of distance. In an inertial frame, in special relativity

$$ds^2 = d\vec{x}^2 - dt^2 = \eta_{\mu\nu}dx^\mu dx^\nu. \tag{17.4}$$

Note here that we have changed the sign of the metric, as we said we would do, from that used earlier in this text. This is the convention of most workers and texts in general relativity and string theory. The coordinate transformation above for the accelerated observer alters the line element. This suggests we consider the generalization:

$$ds^2 = g_{\mu\nu}(x)dx^\mu dx^\nu. \tag{17.5}$$

The *metric tensor*, $g_{\mu\nu}$, will encode the dynamics and physical effects of gravitation. We will know that there is a non-trivial gravitational field when we cannot find coordinates which make $g_{\mu\nu} = \eta_{\mu\nu}$ everywhere.

To develop a dynamical theory, we would like to write invariant actions (which will yield covariant equations). This problem has two parts. We need to couple the fields we already have to the metric in an invariant way. We also require the analog of the field strength for gravity, which will determine the dynamics of $g_{\mu\nu}$ in much the same way as $F_{\mu\nu}$ determines the dynamics of A_μ. This object is the Riemann tensor, $\mathcal{R}^\mu_{\nu\rho\sigma}$. We will see later that the formal analogy can be made very precise, with an object, the spin connection ω_μ, constructed out of the metric tensor, playing the role of A^μ. The close analogy will also be seen when we discuss Kaluza–Klein theories, where higher-dimensional general coordinate transformations become the lower-dimensional gauge transformations.

We first describe how derivatives and $g_{\mu\nu}$ transform under coordinate transformations. Writing

$$x^\mu = x^\mu(x') \tag{17.6}$$

we have

$$\partial'_\mu \phi(x') = \frac{\partial x^\rho}{\partial x^{\mu'}} \partial_\rho \phi(x) = \Lambda_\mu^{\ \rho}(x)\partial_\rho \phi(x). \tag{17.7}$$

An object which transforms like $\partial_\rho \phi$ is said to be a covariant vector. An object which transforms like $\partial_{\rho_1}\phi\partial_{\rho_2}\phi \ldots \partial_{\rho_n}\phi$ is said to be an nth rank covariant tensor; $g_{\mu\nu}$ is an important example of a tensor. We can obtain the transformation law for $g_{\mu\nu}$ from the invariance of the line element:

$$g'_{\mu\nu} dx^{\mu'} dx^{\nu'} = g_{\mu\nu} \frac{\partial x^\mu}{\partial x^{\rho'}} \frac{\partial x^\nu}{\partial x^{\sigma'}} dx^{\rho'} dx^{\sigma'}. \tag{17.8}$$

So

$$g'_{\mu\nu} = g_{\rho\sigma} \frac{\partial x^\rho}{\partial x^{\mu'}} \frac{\partial x^\sigma}{\partial x^{\nu'}}. \tag{17.9}$$

Now dx^μ transforms with the inverse of Λ:

$$dx'^\mu = \frac{\partial x'^\mu}{\partial x^\rho} dx^\rho, \tag{17.10}$$

where dx^μ is said to be a "contravariant" vector. Indices can be raised and lowered with $g_{\mu\nu}$; $g_{\mu\nu}V^\nu$ transforms like a covariant vector, for example.

Another important object is the volume element, d^4x. Under a coordinate transformation,

$$d^4x = \left|\frac{\partial x}{\partial x'}\right| d^4x'. \tag{17.11}$$

The object in between the vertical lines is the Jacobian of the coordinate transformation, $|\det\Lambda|$. The quantity $\sqrt{-\det(g)}$ transforms in exactly the opposite fashion. So

$$\int d^4x\ \sqrt{-\det(g)}, \tag{17.12}$$

the proper four-volume is invariant.

Let's consider a real scalar field, ϕ. The action, before including gravity, is

$$S = \int d^4x \frac{1}{2}[-\partial_\mu\phi\partial_\nu\phi\eta^{\mu\nu} - m^2\phi^2]. \tag{17.13}$$

To make this invariant, we can replace $\eta^{\mu\nu}$ by $g^{\mu\nu}$, and include a factor $\sqrt{\det(-g)}$ along with the d^4x. So

$$S = \int d^4x\sqrt{\det(-g)}\frac{1}{2}[-\partial_\mu\phi\partial_\nu\phi g^{\mu\nu} - m^2\phi^2]. \tag{17.14}$$

The equations of motion should be *covariant*. They must generalize the equation

$$\partial^2\phi = -V'(\phi). \tag{17.15}$$

The first derivative of ϕ, we have seen, transforms as a vector, V_μ, under coordinate transformations, but the second derivative does not transform simply:

$$\begin{aligned}\partial_\mu V_\nu &= \partial_\mu\left(\frac{\partial x^{\rho'}}{\partial x^\nu}V'_\rho\right)\\ &= \frac{\partial x^{\rho'}}{\partial x^\nu}\frac{\partial x^{\sigma'}}{\partial x^\mu}\partial'_\sigma V'_\rho + \frac{\partial^2 x^{\rho'}}{\partial x^\mu\partial x^\nu}V_\rho.\end{aligned} \tag{17.16}$$

To compensate for the extra, inhomogeneous term, we need a covariant derivative, as in gauge theories. Rather than look at the equations of motion directly, however, we can integrate by parts in the scalar field Lagrangian so as to obtain *second* derivatives. This yields:

$$\sqrt{-g}(g^{\mu\nu}\partial_\mu\partial_\nu\phi + \partial_\mu g^{\mu\nu}\partial_\nu\phi) + g^{\mu\nu}\partial_\mu\sqrt{-g}\phi\partial_\nu\phi. \tag{17.17}$$

To bring this into a convenient form, we need a formula for the derivative of a determinant. We can work this out using the same trick we have used repeatedly in

the case of the path integral. Write:

$$\det(M) = \exp(\mathrm{Tr}\ln(M)) \tag{17.18}$$

so

$$\begin{aligned}\det(M+\delta M) &\approx \exp(\mathrm{Tr}\ln(M) + \ln(1+M^{-1}\delta M)) \\ &= \det(M)(1+M^{-1}\delta M).\end{aligned} \tag{17.19}$$

So, for example,

$$\frac{d\det(M)}{dM_{ij}} = M^{-1}_{ij}\det(M). \tag{17.20}$$

Putting this all together, we have the quadratic term in the action for the scalar field:

$$\phi\left(g^{\mu\nu}\partial_\mu\partial_\nu\phi + \partial_\mu g^{\mu\nu}\partial_\nu\phi + g^{\mu\nu}\frac{1}{2}g^{\rho\sigma}\partial_\mu g_{\rho\sigma}\partial_\nu\phi\right). \tag{17.21}$$

Writing this as

$$\phi g^{\mu\nu}D_\mu\partial_\nu\phi, \tag{17.22}$$

we have for the covariant derivative

$$D_\mu V_\nu = \partial_\mu V_\nu - \Gamma^\lambda_{\mu\nu}V_\lambda. \tag{17.23}$$

Here

$$\Gamma^\lambda_{\mu\nu} = \frac{1}{2}g^{\lambda\rho}(\partial_\mu g_{\rho\nu} + \partial_\nu g_{\rho\mu} - \partial_\rho g_{\mu\nu}). \tag{17.24}$$

Note that $\Gamma^\lambda_{\mu\nu}$ is symmetric in μ, ν. The covariant derivative is often denoted by a semicolon and a Greek letter in the subscript or superscript:

$$D_\mu V_\nu = V_{\mu;\nu}. \tag{17.25}$$

The reader can check that

$$\Gamma^\lambda_{\mu\nu} = \Gamma^{\lambda\prime}_{\mu\nu} - \frac{\partial^2 x^\lambda}{\partial x^\mu \partial x^\nu}, \tag{17.26}$$

which just compensates the extra term in the transformation law. Here Γ is known as the *affine connection* (the components of Γ are also sometimes referred to as the Christoffel symbols, and Γ as the Christoffel connection; it is sometimes written as $\left\{ {\mu \atop \nu\ \rho} \right\}$). With this definition:

$$D_\mu V_\nu = \partial_\mu V_\nu - \Gamma^\lambda_{\mu\nu}V_\lambda \tag{17.27}$$

transforms like a tensor with two indices, $V_{\mu\nu}$. Similarly, acting on contravariant vectors:

$$D_\mu V^\nu = \partial_\mu V^\nu + \Gamma^\nu_{\mu\lambda} V^\lambda \tag{17.28}$$

transforms properly. You can also check that $V_{\mu;\nu;\rho}$ transforms as a third-rank covariant tensor, and so on.

To get some practice, and to see how the metric tensor can encode gravity, let's use the covariant derivative to describe the motion of a free particle. In an inertial frame, without gravity,

$$\frac{d^2x^\mu}{d\tau^2} = 0. \tag{17.29}$$

We make this equation covariant by first rewriting it as:

$$\frac{dx^\rho}{d\tau}\frac{\partial}{\partial x^\rho}\left(\frac{dx^\mu}{d\tau}\right). \tag{17.30}$$

We need to replace the derivative of the vector in the last term by a covariant derivative. The covariant version of the equation is:

$$\frac{dx^\rho}{d\tau} D_\rho \left(\frac{\partial x^\mu}{\partial \tau}\right) \tag{17.31}$$

$$= \frac{\partial x^\rho}{\partial \tau}\frac{\partial^2 x^\mu}{\partial x^\rho \partial \tau} + \Gamma^\mu_{\rho\sigma}\frac{\partial x^\sigma}{\partial \tau}\frac{\partial x^\rho}{\partial \tau}. \tag{17.32}$$

So the equation of motion is:

$$\frac{d^2x^\mu}{d\tau^2} + \Gamma^\mu_{\rho\sigma}\frac{\partial x^\sigma}{\partial \tau}\frac{\partial x^\rho}{\partial \tau} = 0. \tag{17.33}$$

This is known as the geodesic equation. Viewed as Euclidean equations, the solutions are geodesics. For a sphere embedded in flat three-dimensional space, for example, the solutions of this equation are easily seen to be great circles. We should be able to recover Newton's equation for a weak gravitational field. For a weak, static gravitational field we might expect

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} \tag{17.34}$$

with $h_{\mu\nu}$ small. Since the gravitational potential in Newton's theory is a scalar, we might further guess:

$$g_{00} = -(1 + 2\phi) \quad g_{ij} = \delta_{ij} \tag{17.35}$$

Then the non-vanishing components of the affine connection are:

$$\begin{aligned}\Gamma^i_{00} &= \frac{1}{2} g^{ij}\left[\partial_0 g_{i0} + \partial_0 g_{0i} - \partial_i g_{00}\right] \\ &= \partial_i \phi\end{aligned} \tag{17.36}$$

and similarly

$$\Gamma^0_{0i} = -\partial_i \phi. \tag{17.37}$$

In the non-relativistic limit, we can replace τ by t, and we have the equation of motion:

$$\frac{d^2 x^i}{dt^2} = -\partial_i \phi. \tag{17.38}$$

17.2 Curvature

Using the covariant derivative, we can construct actions for scalars and gauge fields. Fermions require some additional machinery; we will discuss this towards the end of the chapter. Instead, we turn to the problem of finding an action for the gravitational field itself. In the case of gauge fields, the crucial object was the field strength, $F_{\mu\nu} = [D_\mu, D_\nu]$. For the gravitational field, we will also work with the commutator of covariant derivatives. We write

$$[D_\mu, D_\nu] V_\rho = \mathcal{R}^\sigma_{\rho\mu\nu} V_\sigma, \tag{17.39}$$

where $\mathcal{R}$ is known as the curvature tensor. For a Euclidean space, it measures what we would naturally call the curvature of the space. It is straightforward to work out an expression for $\mathcal{R}$ in terms of the affine connection:

$$\mathcal{R}^\lambda_{\mu\nu\kappa} = \partial_\kappa \Gamma^\lambda_{\mu\nu} - \partial_\nu \Gamma^\lambda_{\mu\kappa} + \Gamma^\eta_{\mu\nu} \Gamma^\lambda_{\kappa\eta} - \Gamma^\eta_{\mu\kappa} \Gamma^\lambda_{\nu\eta}. \tag{17.40}$$

Unlike F, which is first order in derivatives of A, $\mathcal{R}$ is second order in derivatives of g. As a result, the gravitational action will be first order in $\mathcal{R}$.

Note that $\mathcal{R}$ transforms as a tensor under coordinate transformations. It has important symmetry and cyclicity properties. These are most conveniently described by lowering the first index on $\mathcal{R}$:

$$\mathcal{R}_{\lambda\mu\nu\kappa} = \mathcal{R}_{\nu\kappa\lambda\mu} \tag{17.41}$$

$$\mathcal{R}_{\lambda\mu\nu\kappa} = -\mathcal{R}_{\mu\lambda\nu\kappa} = -\mathcal{R}_{\lambda\mu\kappa\nu} = \mathcal{R}_{\mu\lambda\kappa\nu} \tag{17.42}$$

$$\mathcal{R}_{\lambda\mu\nu\kappa} + \mathcal{R}_{\lambda\kappa\mu\nu} + \mathcal{R}_{\lambda\nu\kappa\mu} = 0. \tag{17.43}$$

Starting with $\mathcal{R}$, we can define other tensors. The most important is the Ricci tensor. This has two indices:

$$\mathcal{R}_{\mu\kappa} = g^{\lambda\nu} \mathcal{R}_{\lambda\mu\nu\kappa}. \tag{17.44}$$

$\mathcal{R}$ is a symmetric tensor,

$$\mathcal{R}_{\mu\kappa} = \mathcal{R}_{\kappa\mu}. \tag{17.45}$$

Also very important is the Ricci scalar:

$$\mathcal{R} = g^{\mu\kappa}\mathcal{R}_{\mu\kappa}. \tag{17.46}$$

Note that $\mathcal{R}$ also satisfies an important identity, similar to the Bianchi identity for $F^{\mu\nu}$ (which gives the homogeneous Maxwell equations):

$$\mathcal{R}_{\lambda\mu\nu\kappa;\eta} + \mathcal{R}_{\lambda\mu\eta\nu;\kappa} + \mathcal{R}_{\lambda\mu\kappa\eta;\nu} = 0. \tag{17.47}$$

17.3 The gravitational action

Having introduced, through the Riemann tensor, a description of curvature, we are in a position to write a generally covariant action for the gravitational field. Linear terms in $\mathcal{R}$, as we noted, will be second order in the derivatives of the metric, so they can provide a suitable action. The action must be a scalar, so we take

$$S_{\text{grav}} = \frac{1}{2\kappa^2}\int d^4x\sqrt{-g}\mathcal{R}. \tag{17.48}$$

To obtain the equations of motion, we need to vary the complete action, including the parts involving matter fields, with respect to $g_{\mu\nu}$. We first consider variation of the terms involving the matter fields. The variation of the matter action with respect to $g_{\mu\nu}$ turns out to be nothing but the stress-energy tensor, $T^{\mu\nu}$. Once one knows this fact, this is often the easiest way to find the stress–energy tensor for a system. To see that this identification is correct, we first show that $T_{\mu\nu}$ is covariantly conserved, i.e.

$$D_\nu T^{\nu\mu} = T^{\mu\nu}_{\ \ ;\nu} = 0. \tag{17.49}$$

By assumption, the fields solve the equations of motion in the gravitational background, so the variation of the action, for any variation of the fields, is zero. Consider, then, a space-time translation:

$$x^{\mu'} = x^\mu + \epsilon^\mu. \tag{17.50}$$

Starting with

$$g'_{\mu\nu}(x') = \frac{\partial x^\rho}{\partial x^{\mu'}} g_{\rho\sigma} \frac{\partial x^\sigma}{\partial x^{\nu'}}, \tag{17.51}$$

we have

$$g'_{\mu\nu}(x+\epsilon) = g_{\mu\nu}(x) - \partial_\mu\epsilon^\rho g_{\rho\nu} - \partial_\nu\epsilon^\sigma g_{\sigma\mu}. \tag{17.52}$$

So

$$\delta g_{\mu\nu}(x) = -g_{\mu\lambda}\partial_\nu\epsilon^\lambda - g_{\lambda\nu}\partial_\mu\epsilon^\lambda - \partial_\lambda g_{\mu\nu}\epsilon^\lambda. \tag{17.53}$$

Defining

$$\frac{\delta S_{\rm matt}}{\delta g_{\mu\nu}} = T^{\mu\nu}, \tag{17.54}$$

under this particular variation of the metric we have:

$$\delta S_{\rm matt} = -\int d^4x\sqrt{-g}T^{\mu\nu}[g_{\mu\lambda}\partial_\nu\epsilon^\lambda + g_{\lambda\nu}\partial_\mu\epsilon^\lambda + \partial_\lambda g_{\mu\nu}\epsilon^\lambda]. \tag{17.55}$$

Integrating by parts on the first two terms, and using the symmetry of the metric (and consequently the symmetry of $T^{\mu\nu}$):

$$\delta S_{\rm matt} = \int d^4x[\partial_\mu(T^{\mu\lambda}\sqrt{-g}) - \frac{1}{2}\partial_\lambda g_{\mu\nu}T^{\mu\nu}\sqrt{-g}]\epsilon^\lambda. \tag{17.56}$$

The coefficient of ϵ^λ vanishes for fields which obey the equations of motion. This object is $T^{\mu\nu}_{\ \ ;\mu}$. The reader can verify this last identification painstakingly, or by noting that:

$$\Gamma^\mu_{\mu\lambda} = \frac{1}{\sqrt{-g}}\partial_\lambda\sqrt{-g} \tag{17.57}$$

so, for a general vector, for example,

$$V^\mu_{\ ;\mu} = \frac{1}{\sqrt{-g}}\partial_\mu(\sqrt{-g}V^\mu), \tag{17.58}$$

and similarly for higher-rank tensors.

As a check, consider the stress tensor for a free massive scalar field. Once more, the action is:

$$S = \int d^4x\sqrt{-g}\left(-\frac{1}{2}g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi - \frac{1}{2}m^2\phi^2\right). \tag{17.59}$$

So, recalling our formula for the variation of the determinant:

$$T_{\mu\nu} = \frac{1}{2}\partial_\mu\phi\partial_\nu\phi - \frac{1}{4}g_{\mu\nu}(g^{\rho\sigma}\partial_\rho\phi\partial_\sigma\phi - m^2\phi^2). \tag{17.60}$$

To find the full gravitational equations – Einstein's equations – we need to vary also the gravitational term in the action. This is best done by explicitly constructing the variation of the curvature tensor under a small variation of the field. We leave the details for the exercises, and quote the final result:

$$\mathcal{R}_{\mu\nu} - \frac{1}{2}g_{\mu\nu}\mathcal{R} = \kappa^2 T_{\mu\nu}. \tag{17.61}$$

We will consider many features of this equation, but it is instructive to see how we obtain Newton's expression for the gravitational field, in the limit that gravity is

not too strong. We have already argued that in this case we can write

$$g_{00} = -(1 + 2\phi) \quad g^{ij} = \delta_{ij}. \tag{17.62}$$

As we have seen, the non-vanishing components of the connection are:

$$\Gamma^i_{00} = \partial_i \phi \quad \Gamma^0_{i0} = -\partial_i \phi. \tag{17.63}$$

Correspondingly, the non-zero components of the curvature tensor are:

$$\mathcal{R}^i_{00j} = \partial_i \partial_j \phi = -\mathcal{R}^i_{0j0} = \mathcal{R}^0_{ij0} \tag{17.64}$$

where the relations between the various components follow from the symmetries of the curvature tensor. From these we can construct the Ricci tensor and the Ricci scalar:

$$\mathcal{R}_{00} = \nabla^2 \phi \quad \mathcal{R} = -\nabla^2 \phi. \tag{17.65}$$

So we obtain

$$-\nabla^2 \phi = \kappa^2 T_{00}. \tag{17.66}$$

Note, from this, we can identify Newton's constant in terms of κ,

$$G_{\mathrm{N}} = \frac{\kappa^2}{8\pi}. \tag{17.67}$$

17.4 The Schwarzschild solution

Not long after Einstein wrote down his equations for general relativity, Schwarzschild constructed the solution of the equations for a static, isotropic metric. Such a metric can be taken to have the form:

$$ds^2 = -B(r)dt^2 + A(r)dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2). \tag{17.68}$$

Actually, rotational invariance would allow other terms. In terms of vectors, $d\vec{x}$, the most general metric has the form

$$-B(r)dt^2 + D(r)\vec{x} \cdot d\vec{x}dt + C(r)d\vec{x} \cdot d\vec{x} + D(r)(\vec{x} \cdot d\vec{x})^2. \tag{17.69}$$

By a sequence of coordinate transformations, however, one can bring the metric to the form above.

We will solve Einstein's equations with $T_{\mu\nu} = 0$. Corresponding to ds^2, we have the non-vanishing metric components:

$$g_{rr} = A(r) \quad g_{\phi\phi} = r^2 \sin^2\theta \quad g_{tt} = -B(r) \quad g_{\theta\theta} = r^2. \tag{17.70}$$

Our goal is to determine A and B. The equations for A and B follow from Einstein's equations. We first need to evaluate the non-vanishing Christoffel symbols. This is done in the exercises. While straightforward, the calculation of the connection and the curvature is slightly tedious, and this is an opportunity to practice with the computer packages described in the exercises. The non-vanishing components of the affine connection are:

$$\Gamma^r_{rr} = \frac{1}{2A(r)}A'(r) \quad \Gamma^r_{\theta\theta} = -\frac{r}{A(r)} \quad \Gamma^r_{\phi\phi} = -\frac{r\sin^2(\theta)}{A(r)}$$
$$\Gamma^r_{\phi\phi} = \frac{r\sin^2\theta}{A(r)} \quad \Gamma^r_{tt} = \frac{1}{2A(r)}B'(r), \tag{17.71}$$

where the primes denote derivatives with respect to r. Similarly

$$\Gamma^\theta_{r\phi} = \Gamma^\theta_{\theta r} = \frac{1}{r} \quad \Gamma^\theta_{\phi\phi} = -\sin(\theta)\cos(\theta)$$
$$\Gamma^\phi_{\phi r} = \Gamma^\phi_{r\phi} = \frac{1}{r} \quad \Gamma^\phi_{\phi\theta} = \Gamma^\phi_{\theta\phi} = \cos(\theta)$$
$$\Gamma^t_{tr} = \Gamma^t_{rt} = \frac{B'}{2B}. \tag{17.72}$$

The non-vanishing components of the Ricci tensor are:

$$R_{rr} = \frac{B''}{2B} - \frac{1}{4}\frac{B''}{B}\left(\frac{A'}{A} + \frac{B'}{B}\right) - \frac{1}{r}\frac{A'}{A} \tag{17.73}$$

$$R_{\theta\theta} = -1 + \frac{r}{2A}\left(-\frac{A'}{A} + \frac{B'}{B}\right) + \frac{1}{A} \tag{17.74}$$

$$R_{\phi\phi} = \sin^2\theta R_{\theta\theta} \quad R_{tt} = -\frac{B''}{2A} + \frac{1}{4}\frac{B'}{A}\left(\frac{A'}{A} + \frac{B'}{B}\right) - \frac{1}{r}\frac{B'}{A}. \tag{17.75}$$

For empty space, Einstein's equation reduces to

$$R_{\mu\nu} = 0. \tag{17.76}$$

We will require that asymptotically, the space-time is just flat Minkowski space, so we will solve these equations with the requirement

$$A_{r\to\infty} = B_{r\to\infty} = 1. \tag{17.77}$$

Examining the components of the Ricci tensor, we see that it is enough to set $R_{rr} = R_{\theta\theta} = R_{tt} = 0$. We can simplify the equations with a little cleverness:

$$\frac{R_{rr}}{A} + \frac{R_{tt}}{B} = -\frac{1}{rA}\left(\frac{A'}{A} + \frac{B'}{B}\right). \tag{17.78}$$

From this it follows that $A = 1/B$. Then, from $R_{\theta\theta} = 0$, we have

$$\frac{d}{dr}(rB) - 1 = 0. \tag{17.79}$$

So

$$rB = r + \text{ const.} \tag{17.80}$$

Now $B = -g_{tt}$, so, far away, where the space-time is nearly flat, $B = 1 + 2\phi$, where ϕ is the gravitational potential. So:

$$B(r) = \left[1 - \frac{2MG}{r}\right] \quad A(r) = \left[1 - \frac{2MG}{r}\right]^{-1}. \tag{17.81}$$

17.5 Features of the Schwarzschild metric

So finally, we have the Schwarzschild metric:

$$ds^2 = -\left[1 - \frac{2MG}{r}\right]dt^2 + \left[1 - \frac{2MG}{r}\right]^{-1} dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2. \tag{17.82}$$

Far away, this clearly describes an object of mass M. While we have so far discussed the energy–momentum tensor for matter, we have not discussed the energy of gravitation. The situation is similar to the problem of defining charge in a gauge theory. There, the most straightforward definition involves using the asymptotic behavior of the fields to determine the total charge. In gravity, the energy is similar. There is no invariant local definition of the energy density. But in spaces that are asymptotically flat, one can give a global notion of the energy, known as the ADM (for Arnowitt, Deser and Misner) energy. Only the $1/r$ behavior of the fields enters. We will not review this here, but, not surprisingly, in the present case, this energy, P^0, is equal to M.

The curvature of space-time near a star yields observable effects. Einstein, when he first published his theory, proposed two tests of the theory: the bending of light by the Sun, and the precession of Mercury's perihelion. In the latter case, the theory accounted for a known anomaly in the motion of the planet; the prediction of the bending of light was soon measured.

A striking feature of this metric is that it becomes singular at a particular value of r, known as the Schwarzschild radius (horizon),

$$r_{\rm h} = 2MG. \tag{17.83}$$

At this point, the coefficient of dr^2 diverges, and that of dt^2 vanishes. Both change sign, in some sense reversing the roles of r and t. This singularity is a bit of a fake. No component of the curvature becomes singular. One can exhibit this by choosing

coordinates in which the metric is completely non-singular (see the exercises at the end of the chapter).

For most realistic objects, such as planets and typical stars, this r_h is well within the star, where surely it is important to use a more realistic model of $T_{\mu\nu}$. But there are systems in nature where the "material" lies well within the Schwarzschild radius. These systems are known as *black holes.* The known black holes arise from the collapse of very massive stars. It is conceivable that smaller black holes arise from more microscopic processes. These systems have striking properties. Classically, light cannot escape from within the horizon; the curvature singularity at the origin is real. Black holes are nearly featureless. Classically, an external observer can only determine the mass, charge and angular momentum of the black hole, however complex the system which may have preceded it.

Bekenstein pointed out that the horizon area has peculiar properties, and behaves much like a thermal system. Most importantly, it obeys a relation analogous to the second law,

$$dA > 0. \tag{17.84}$$

Identifying the area with an entropy suggests that one can associate a temperature with the black hole, known as the Hawking temperature. The black hole horizon is a sphere of area $4\pi r_h^2$. So one might guess, on dimensional grounds,

$$T_h = \frac{1}{8\pi GM}. \tag{17.85}$$

The precise constant does not follow from this argument. The reader is invited to work through a heuristic, path integral derivation in the exercises.

Quantum mechanically, Hawking showed that this temperature has a microscopic significance. When one studies quantum fields in the gravitational background, one finds that particles escape from the black hole. These particles have a thermal spectrum, with a characteristic temperature T_h.

These features of black holes raise a number of conceptual questions. For the black hole at the center of the galaxy, for example, with mass millions of times greater than the Sun, the Hawking temperature is ludicrously small. Correspondingly, the Hawking radiation is totally irrelevant. But one can imagine microscopic black holes which would evaporate in much more modest periods of time. This raises a puzzle. The Hawking radiation is strictly thermal. So one could form a black hole, say, in the collapse of a small star. The initial star is a complex system, with many features. The black hole is nearly featureless. Classically, however, one might imagine that some memory of the initial state of the system is hidden behind the horizon; this information would simply be inaccessible to the external observer. But owing to the evaporation, the black hole, and its horizon, eventually

disappear. One is left with just a thermal bath of radiation, with features seemingly determined by the temperature (and therefore the mass). Hawking suggested that this information paradox posed a fundamental challenge for quantum mechanics: pure states could evolve into mixed states, through the formation of a black hole. For many years, this question was the subject of serious debate. One might respond to Hawking's suggestion by saying that the information is hidden in subtle correlations in the radiation, as would be the case of burning, say, a lump of coal initially in a pure state. But more careful consideration indicates that things cannot be quite so simple. Only in relatively recent years has string theory provided at least a partial resolution of this paradox. We will touch on this subject briefly in the chapters on string theory. In the suggested reading, the reader will be referred to more thorough treatments.

17.6 Coupling spinors to gravity

In any theory which will ultimately describe nature, both spinors and general relativity will be present. Coupling spinors to gravity requires some concepts beyond those we have utilized up to now. The usual covariant derivative is constructed for tensors under changes of coordinates. In flat space, spinors are defined by their properties under rotations – more generally Lorentz transformations. To do the same in general relativity, it is necessary, first, to introduce a local Lorentz frame at each point. The basis vectors in this frame are denoted e^a_μ. Here μ is the Lorentz index; we can think of a as labeling the different vectors. The e_μs, in four dimensions, are referred to as a tetrad or "vierbein." In other dimensions, these are called vielbein.

Requiring that the basis vectors be orthonormal, in the Lorentzian sense, gives

$$e^a_\mu(x)e_{a\nu}(x) = g_{\mu\nu}(x) \tag{17.86}$$

or, equivalently,

$$e^a_\mu(x)e^{b\mu}(x) = \eta^{ab}. \tag{17.87}$$

The choice of the vielbein is not unique. We can multiply e by a Lorentz matrix, $\Lambda^a_b(x)$. Using e, we can change indices from space-time (sometimes called "world") indices to tangent space indices:

$$V^a = e^a_\mu V^\mu. \tag{17.88}$$

Using this, we can figure out what is the form of the connection which maintains the gauge symmetry. We require:

$$D_\mu V^a = e^{a\nu} D_\mu V_\nu. \tag{17.89}$$

The derivative on the left-hand side is:

$$\left(\partial_\mu V^a + (\omega_\mu)^a_b V^b\right) \tag{17.90}$$

With a bit of work, one can find explicitly the connection between the spin connection and the vielbein:

$$\omega^{ab}_\mu = \frac{1}{2}e^{\nu a}\left(\partial_\mu e^b_\nu - \partial_\nu e^b_\mu\right) - \frac{1}{2}e^{\nu b}\left(\partial_\mu e^a_\nu - \partial_\nu e^a_\mu\right) - \frac{1}{2}e^{\rho a}e^{\sigma b}(\partial_\rho e_{\sigma c} - \partial_\sigma e_{\rho c})e^c_\mu. \tag{17.91}$$

Now we put this together. First, the curvature has a simple expression in terms of the spin connection, which formally is identical to that of a Yang–Mills connection:

$$(R_{\mu\nu})^a_b = \partial_\mu(\omega_\nu)^a_b - \partial_\nu(\omega_\mu)^a_b + [\omega_\mu, \omega_\nu]^a_b. \tag{17.92}$$

This is connected simply to the Riemann tensor:

$$(R_{\mu\nu})^a_b = e^a_\sigma e^\tau_b (R_{\mu\nu})^\sigma_\tau. \tag{17.93}$$

We can now construct, also, a generally covariant action for spinors:

$$\int d^D x \sqrt{g} i \bar{\psi}\Gamma^a e^\mu_a \left(\partial_\mu + \frac{1}{2}\omega^{bc}_\mu \Sigma_{bc}\right)\psi. \tag{17.94}$$

Suggested reading

There are a number of excellent textbooks on general relativity, for example those of Weinberg (1972), Wald (1984), Carroll (2004) and Hartle (2003). Many aspects of general relativity important for string theory are discussed in the text of Green *et al.* (1987). A review of black holes in string theory is provided by Peet (2000).

Exercises

(1) Show that $g^{\mu\nu}\partial_\nu$ transforms like dx^μ. Verify explicitly that the covariant derivative of a vector transforms correctly.

(2) Derive Eq. (17.38) by considering the action for a particle:

$$S = -\int ds = -\int \sqrt{-g_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau}}. \tag{17.95}$$

(3) Verify the formula, Eq. (17.40) for $\mathcal{R}$, its symmetry properties, and the Bianchi identities.

(4) Repeat the derivation of the conservation of the stress tensor, being careful with all of the steps. Derive the stress tensor for the Maxwell field of electrodynamics, $F_{\mu\nu}$. Derive Einstein's equations from the action. You'll want to show first that

$$\delta R_{\mu\nu} = \left(\delta\Gamma^\lambda_{\mu\lambda}\right)_{;\nu} - \left(\delta\Gamma^\lambda_{\mu\nu}\right)_{;\lambda}.$$

(5) Download a package of programs for doing calculations in general relativity in MAPLE, MATHEMATICA, or any other program you prefer. A Google search will yields several choices. Practice by computing the components of the affine connection and the curvature for the Schwarzschild solution.

(6) Heuristic derivation of the Hawking temperature: near the horizon, one can choose coordinates so that the metric is almost flat. Check that with

$$\eta = 2\sqrt{r_{\rm h}(r - r_{\rm h})} \tag{17.96}$$

$$ds^2 = -4r_{\rm h}^2\eta^2 dt^2 + d\eta^2 + r_{\rm h}^2 d\Omega_2^2. \tag{17.97}$$

Now take the time to be Euclidean, $t \to i\phi/(2r_{\rm h})$. Check that now this is of the metric of the plane times that of a two sphere, provided that ϕ is an angle, $0 < \phi < 2\pi$ (otherwise, the space is said to have a conical singularity). Argue that field theory on this sphere is equivalent to field theory at finite temperature, with temperature $T_{\rm h}$ (you may need to read Appendix C, particularly the discussion of finite temperature field theory).

18

Cosmology

Very quickly after Einstein published his general theory, a number of researchers attempted to apply Einstein's equations to the universe as a whole. This was a natural, if quite radical, move. In Einstein's theory, the distribution of energy and momentum in the universe determines the structure of space-time, and this applies as much to the universe as a whole as to the region of space, say, around a star. To get started these early researchers made an assumption which, while logical, may seem a bit bizarre. They took the principles enunciated by Copernicus to their logical extreme, and assumed that space-time was homogeneous and isotropic, i.e. that there is no special place or direction in the universe. They had virtually no evidence for this hypothesis at the time – definitive observations of galaxies outside of the Milky Way were only made a few years later. It was only decades later that evidence in support of this *cosmological principle* emerged. As we will discuss, we now know that the universe is extremely homogeneous, when viewed on sufficiently large scales.

To implement the principle, just as, for the Schwarzschild solution, we begin by writing the most general metric consistent with an assumed set of symmetries. In this case, the symmetries are homogeneity and isotropy in space. A metric of this form is called *Friedmann–Robertson–Walker* (FRW). We can derive this metric by imagining our three-dimensional space, at any instant, as a surface in a four-dimensional space. There should be no preferred direction on this surface; in this way, we will impose both homogeneity and isotropy. The surface will then be one of constant curvature. Consider, first, the mathematics required to describe a $2+1$-dimensional space-time of this sort. The three *spatial* coordinates would satisfy

$$x_1^2 + x_2^2 = k\big(R^2 - X_3^2\big), \tag{18.1}$$

where $k = \pm 1$ might be positive or negative, corresponding to a space of positive

or negative curvature. Then the line element on the surface is (for positive k):

$$d\vec{x}^2 = dx_1^2 + dx_2^2 + dx_3^2 = dx_1^2 + dx_2^2 + \frac{(x_1 dx_1 + x_2 dx_2)^2}{x_3^2}. \tag{18.2}$$

The equation of the hypersurface gives

$$x_3^2 = R^2 - x_1^2 - x_2^2. \tag{18.3}$$

Calling $x_1 = r' \cos(\theta)$, $x_2 = r' \sin(\theta)$,

$$d\vec{x}^2 = \frac{R^2 dr'^2}{R^2 - r'^2} + r'^2 d\theta^2. \tag{18.4}$$

It is natural to rescale $r' = r/R$. Then the metric takes the form, now for general k:

$$d\vec{x}^2 = \frac{dr'^2}{1 - kr^2} + r^2 d\theta^2. \tag{18.5}$$

Here $k = 1$ for a space of positive curvature; $k = -1$ for a space of negative curvature; $k = 0$ is a spacial case, corresponding to a flat universe.

We can immediately generalize this to three dimensions, allowing the radius, R, to be a function of time, $R \to a(t)$. In this way we obtain the Friedman–Robertson–Walker (FRW) metric:

$$ds^2 = -dt^2 + a^2(t) \left\{ \frac{dr^2}{1 - kr^2} + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2 \right\}. \tag{18.6}$$

By general coordinate transformations, this can be written in a number of other convenient and commonly used forms, which we will encounter in the following.

First, let's evaluate the connection and the curvature. Again, the reader should evaluate a few of these terms by hand, and perform the complete calculation using one of the programs mentioned in the exercises in the previous chapter. The non-vanishing components of the connection are:

$$\Gamma^i_{0j} = \frac{\dot{a}}{a}\delta^i_j \quad \Gamma^0_{ij} = g_{ij}\frac{\dot{a}}{a} \quad \Gamma^i_{jk} = \frac{g^{il}}{2}[g_{lj,k} + g_{lk,j} - g_{jk,l}] \tag{18.7}$$

and of the curvature are:

$$\begin{aligned} \mathcal{R}_{00} &= -3\frac{\ddot{a}}{a} \\ \mathcal{R}_{ij} &= g_{ij}\left[\frac{\ddot{a}}{a} + 2H^2 - 2\frac{k}{a^2}\right]. \end{aligned} \tag{18.8}$$

Here H is known as the Hubble parameter,

$$H = \frac{\dot{a}}{a}, \tag{18.9}$$

and represents the expansion rate of the universe. Today

$$H = 100\,h\ \text{km s}^{-1}\ \text{Mpc}^{-1} \quad h = 0.73 \pm 0.03. \tag{18.10}$$

The assumption of homogeneity and isotropy greatly restricts the form of the stress tensor; $T_{\mu\nu}$ must take the *perfect fluid* form:

$$T_{00} = \rho \quad T_{ij} = p g_{ij} \tag{18.11}$$

where ρ and p are the energy density and the pressure, and are assumed to be functions only of time. Then the $0-0$ component of the Einstein equation gives the *Friedmann equation*:

$$\frac{\dot{a}^2}{a^2} + \frac{k}{a^2} = \frac{8\pi G}{3}\rho. \tag{18.12}$$

The $i - j$ components give:

$$\frac{2\ddot{a}}{a} + \frac{\dot{a}^2}{a^2} + \frac{k}{a^2} = -8\pi G\rho. \tag{18.13}$$

There is also an equation which follows from the conservation of the energy momentum tensor, $T^{\mu\nu}_{\ ;\nu} = 0$. This is

$$d(\rho a^3) = -p d(a^3). \tag{18.14}$$

This equation is familiar in thermodynamics as the equation of energy conservation, if we interpret a^3 as the volume. Suppose that we have the equation of state $p = w\rho$, where w is a constant. Then Eq. (18.14) says:

$$\rho \propto a^{-3(1+w)}. \tag{18.15}$$

Three special cases are particularly interesting. For non-relativistic matter, the pressure is negligible compared to the energy density, so $w = 0$. For radiation (relativistic matter), $w = 1/3$. For a Lorentz-invariant stress tensor, $T_{\mu\nu} = \Lambda g_{\mu\nu}$, $p = -\rho$, so $w = -1$. For these cases, it is worth remembering that

$$\text{Radiation:}\rho \propto R^{-4} \quad \text{Matter:}\rho \propto R^{-3}; \quad \text{Vacuum:}\rho = \text{constant}. \tag{18.16}$$

After taking account conservation of stress-energy and the Bianchi identities, only one of the two Einstein equations we have written is independent, and it is conventional to take the Friedmann equation. This equation can be rewritten in terms of the Hubble parameter,

$$\frac{k}{H^2 a^2} = \frac{8\pi G\rho}{3H^2} - 1. \tag{18.17}$$

Examining the right-hand side of this equation, it is natural to define a *critical density*,

$$\rho_c = \frac{3H^2}{8\pi G}, \tag{18.18}$$

and to define Ω as the ratio of the density to the critical density,

$$\Omega = \frac{\rho}{\rho_c}. \tag{18.19}$$

So $k = 1$ corresponds to $\Omega > 1$, $k = -1$ to $\Omega < 1$, and $k = 0$, a flat universe, to $\Omega = 1$. It is also natural to break up Ω into various components, such as radiation, matter, and cosmological constant. As we will discuss shortly, Ω today is equal to one within experimental errors; its main components are baryons, some unknown form of matter, and dark energy (perhaps cosmological constant),

$$\Omega_b = 0.04; \quad \Omega_{dm} = 0.28; \quad \Omega_{de} = 0.73. \tag{18.20}$$

The present error bars are of order 4% on Ω_b, and 13% on the dark matter density. The errors are similar on the dark energy determination.

The history of the universe divides into various eras, in which different forms of energy were dominant. The earliest era for which we have direct observational evidence is a period lasting from a few seconds after the big bang to about 100 000 years, during which the universe was radiation dominated. From the Friedmann equation, setting $k = 0$, we have that

$$a(t) = a(t_0)t^{1/2} \quad H = \frac{1}{2t}. \tag{18.21}$$

For the period of matter domination, which began about 10^5 years after the big bang and lasted almost to the present:

$$a(t) \propto t^{2/3} \quad H = \frac{2}{3t}. \tag{18.22}$$

The universe appears today to be passing from an era of matter domination to a phase in which a (positive) cosmological constant dominates. Such a space is called a de Sitter space, and:

$$a(t) \propto e^{H_d t} \quad H_d = \frac{8\pi G}{3}\Lambda. \tag{18.23}$$

In the radiation dominated and matter dominated periods, H is, as we see from the formulas above, roughly a measure of the age of the universe. One can define the age of the universe more formally as:

$$t = \int^{a(t)} \frac{da}{\dot{a}} = \int \frac{da}{aH}. \tag{18.24}$$

The present value of the Hubble constant corresponds to $t \approx 15$ billion years. To obtain this correspondence between the age and the measured H_0, it is important to include both the matter and the cosmological constant pieces in the energy density. Note, in particular, that the integral is dominated by the most recent times, where H is smallest.

18.1 A history of the universe

As little as 50 years ago, most scientists would have been surprised at just how much we would eventually know about the universe: its present composition, its age, and its history, back to times a couple of minutes after the big bang. We have direct evidence of phenomena at much earlier times, though its full implications are difficult to interpret. We understand how galaxies formed and the abundance of the light elements. And we have a trove of plausible speculations, some of which we should be able to test over time.

In this section, we outline some of the basic features of this picture. Examining the FRW solution of Einstein's equations, we see that the scale factor, $a(t)$, gets monotonically smaller in the past. The Hubble parameter, H, becomes larger. So at some time, the universe was much smaller than it is today. More precisely, the objects we see, or their predecessors, were far closer together. Far enough back in time, the material we currently see was highly compressed, and hot. So at some stage, the universe was likely dominated by radiation. Recall that, during a radiation-dominated era,

$$a \sim t^{1/2} \quad H = \frac{1}{2t}. \tag{18.25}$$

If we suppose that the universe remained radiation dominated as we look further back in time, we face a problem. At $t = 0$, the metric is singular – the curvature diverges. This is a finite time in the past, since

$$\int_0^{\text{today}} dt \sqrt{-g_{00}} \tag{18.26}$$

converges as $t \to 0$. This is not simply a feature of our particular assumptions about the equation of state or the precise form of the metric, but a feature of solution of Einstein's equations; it is a consequence of singularity theorems due to Penrose and Hawking. The meaning of this singularity is a subject of much speculation. It might be smoothed out by quantum effects, or indicate something else. For now, we simply have to accept that some early time is inaccessible to us. To make progress, we suppose, first, that at time t_0, the universe was extremely hot, with temperature T_0, and reasonably homogeneous and isotropic. We then allow the universe to

evolve, using Einstein's equations, the known particles and their interactions, and basic principles of statistical mechanics. As we will see, we can safely take T_0 at least as large as several MeV (corresponding to temperatures larger than 10^{10} K).

To make further progress, we need to think about the content of the universe and how it evolves as the universe expands. The universe cannot be precisely in thermal equilibrium, but for much of its history it is very nearly so, with matter and radiation evolving adiabatically. To understand why the expansion is adiabatic, note first that H^{-1} is a time-scale for the expansion. If the universe is radiation dominated,

$$H \sim \frac{T^2}{M_{\rm p}}. \tag{18.27}$$

The rate for interactions in a gas will scale with T, multiplied, perhaps, by a few powers of coupling constants. For temperatures well below the Planck scale, the reaction rates will be much more rapid than the expansion rate. So at any given instant, the system will nearly be in equilibrium.

It is worth reviewing a few formulas from statistical mechanics. These formulas can be derived by elementary considerations, or by using the methods of finite-temperature field theory, as discussed in Appendix C. For a relativistic, weakly coupled Bose gas,

$$\rho = \frac{\pi^2}{30} g T^4 \quad p = \frac{\rho}{3}, \tag{18.28}$$

while for a similar Fermi gas:

$$\rho = \frac{7}{8}\frac{\pi^2}{30} g T^4 \quad p = \frac{\rho}{3}. \tag{18.29}$$

Here g is a degeneracy factor, counting the number of physical helicity states of each particle type. In the non-relativistic limit, for both bosons and fermions,

$$n = g\left(\frac{mT}{2\pi}\right)^{3/2} \exp(-(m-\mu)/T) \tag{18.30}$$

$$\rho = mn \quad p = nT \ll \rho. \tag{18.31}$$

For temperatures well below m, the density rapidly goes to zero unless $\mu \neq 0$. Note that μ may be non-zero when there is a (possibly approximately) conserved quantum number. Perhaps the most notable example is baryon number.

We should pause here and discuss an aspect of general relativity which we have not considered up to now. A gravitational field alters the behavior of clocks. This is known as the gravitational red-shift. We can understand this in a variety of ways. First, if we have a particle at rest in a gravitational field, we see that the proper time is related to the coordinate time by a factor $\sqrt{g_{00}}$. Consider, alternatively,

the equation for a massless scalar field with momentum k in an expanding FRW universe. This is just $D^\mu \partial_\mu \phi = 0$. Using the non-vanishing Christoffel symbols, with $\phi(\vec{x}, t) = e^{i\vec{k}\cdot\vec{x}}\phi(t)$,

$$\ddot{\phi}(k) + 3H\dot{\phi}(k) + \frac{k^2}{a^2(t)}\phi = 0. \tag{18.32}$$

As a result of the last term, the wavelength effectively decreases as $1/a(t)$. A photon red-shifts in precisely the same way.

The implications of this for the statistical mechanical distribution functions are interesting. Consider, first, a massless particle such as the photon. For such a particle, the distribution is:

$$\int \frac{d^3k}{(2\pi)^3} \frac{1}{e^{k/T} - 1}. \tag{18.33}$$

The effect of the red-shift is to maintain this form of the distribution, but to change the temperature, $T(t) \propto 1/a(t)$. So even if the particles are not in equilibrium, they maintain an equilibrium distribution appropriate to the red-shifted temperature. This is not the case for massive particles.

Let's imagine, then, starting the clock at temperatures well above the scale of QCD, but well below the scale of weak interactions, say at 10 GeV. In this regime, the density of Ws and Zs is negligible, but the quarks and gluons behave as nearly free particles. So we can take an inventory of bosons and fermions light compared to T. The bosons include the photon and the gluons; the fermions include all of the quarks and leptons except the top quark. So the effective g, which we might call g_{10}, is approximately 98. This means, for example, that

$$\rho \approx \frac{g_{10}\pi^2}{30} T^4 \tag{18.34}$$

and the Hubble constant is related to the temperature through:

$$H = \left[\frac{8\pi}{3} G \frac{\pi^2}{30} g_{10} T^4\right]^{1/2}. \tag{18.35}$$

This allows us to write a precise formula for the temperature as a function of time:

$$T = \left[\frac{16\pi}{3} G \frac{\pi^2}{30} g_{10}\right]^{-1/4} \left(\frac{1}{t}\right)^{1/2}. \tag{18.36}$$

As the universe cools, QCD changes from a phase of nearly free quarks and gluons to a hadronic phase. At temperatures below m_π, the only light species are the electron and the neutrinos. By this time, the anti-neutrons have annihilated with neutrons and the anti-protons against protons, leaving a small net baryon number,

the total number of neutrons and protons. There is, at this time, of order one baryon per billion photons. We will have much more to say about this slight excess later.

At this stage, interactions involving neutrinos maintain an equilibrium distribution of protons and neutrons. We can give a crude, but reasonably accurate, estimate of the temperature at which neutrino interactions drop out of equilibrium by asking when the interaction rate becomes comparable to the expansion rate. The cross section for neutrinos off of protons behaves as:

$$\sigma(\nu + p \to n + e) \approx G_{\rm F}^2 E^2 \tag{18.37}$$

and the number density of neutrinos is

$$n_\nu \approx \frac{\pi^2}{30} g_T T^3. \tag{18.38}$$

Combining this with our formula for the Hubble constant as a function of T, Eq. (18.35), gives for the decoupling temperature, T_ν,

$$T_\nu^3 \approx G_{\rm F}^{-2} M_{\rm p}^{-1}, \tag{18.39}$$

or

$$T_\nu \approx 2\ {\rm MeV}. \tag{18.40}$$

This corresponds to a time of order 100 s after the big bang. At this point, neutron decays are not compensated by the inverse reaction. On the other hand, many neutrons will pair with protons to form stable light elements such as D and He. At about this time, the abundances of the various light elements are fixed.

There is a long history of careful, detailed calculations of the abundances of the light elements (H, He, D, Li, $\cdots$) which result from this period of decoupling. The abundances turn out to be a sensitive function of the ratio of baryon to photons, $n_{\rm B}/n_\gamma$. Astronomers have also made extensive efforts to *measure* this ratio. A comparison of observations and measurements. gives for the baryon to photon ratio:

$$\frac{n_{\rm B}}{n_\gamma} = 6.1^{+0.3}_{-0.2} \times 10^{-10}. \tag{18.41}$$

We will see that this result receives strong corroboration from other sources.

The universe continues to cool in this radiation dominated phase for a long time. At $t \approx 10^5$ years, the temperature drops to about 1 eV. At this time, electrons and nuclei can combine to form neutral atoms. As the density of ionized material drops, the universe becomes essentially transparent to photons. This is referred to as *recombination.* The photons now stream freely. As the universe continues to cool, the photons red-shift, maintaining a Planck spectrum. Today, these photons behave as if they had a temperature $T \approx 3$ K. They constitute the *cosmic microwave*

background radiation (CMBR). This radiation was first observed in 1963 by Penzias and Wilson, and has since been extensively studied. It is very precisely a black body, with characteristic temperature 2.7 K. We will discuss other features of this radiation shortly.

It is interesting that, given the measured value of the matter density, matter and radiation have comparable energy densities at recombination time. At later times, matter dominates the energy density, and this continues to be the case to the present time.

In our brief history, another important event occurs at $t \simeq 10^9$ years. If we suppose that initially there were small inhomogeneities, these remain essentially frozen, as we will explain later, until the time of matter-radiation equality. They then grow with time. From observations of the CMBR, we know that these inhomogeneities were at the level of one part in 10^5. At about 1 billion years after the big bang, these then grow enough as to be non-linear, and their subsequent evolution is believed to give rise to the structure – galaxies, clusters of galaxies, and so on, that we see around us.

One surprising feature of the universe is that most of the energy density is in two forms which we cannot see directly. One is referred to as the "dark matter." The possibility of dark matter was first noted by astronomers in the 1930s, from observations of rotation curves of galaxies. Simply using Newton's laws, one can calculate the expected rotational velocities and one finds that these do not agree with the observed distribution of stars and dust in the galaxies. This is true for structures on many scales, not only galaxies but clusters and larger structures. Other features of the evolution of the universe are not in agreement with observation unless most of the energy density is in some other form. From a variety of measurements, $\Omega_{\rm m}$, the fraction of the critical energy density in matter, is known to be about 0.3. Nucleosynthesis and the CMBR give a much smaller fraction in baryons, $\Omega_{\rm b} \approx 0.05$. In support of this picture, direct searches for hidden baryons are compatible with the smaller number, and have failed to find anything like the required density to give $\Omega_{\rm m}$.

Finally, it appears that we are now entering a new era in the history of the universe. For the last 14 billion years, the energy density has been dominated by non-relativistic matter. But at the present time, there is almost twice as much energy in some new form, with $p < 0$, referred to as dark energy. This is quite possibly a cosmological constant, Λ. Current measurements are compatible with $w = -1$ ($p = -\rho$).

The picture we have described has extensive observational support. We have indicated some of this: the light element abundances and the observation of the CMBR. The agreement of these two quite different sets of observations for the baryon to photon ratio is extremely impressive. Observations of supernovae, the

age of the universe, and features of structure at different scales all support the existence of a cosmological constant (dark energy) constituting about 70% of the total energy.

This is not a book on cosmology, and the overview we have presented is admittedly sketchy. There are many aspects of this picture we have not discussed. Fortunately, there are many excellent books on the subject, some of which are mentioned as suggested reading.

Suggested reading

There are a number of good books and lectures on aspects of cosmology discussed here. Apart from the text of Weinberg (1972), mentioned earlier, these include the texts of Kolb and Turner (1990) and of Dodelson (2004).

Exercises

(1) Compute the Christoffel symbols and the curvature for the FRW metric. Verify the Friedmann equations.
(2) Verify Eq. (18.32).
(3) Consider the case of de Sitter space, $T_{\mu\nu} = -\Lambda g_{\mu\nu}$, with positive Λ. Show that the space expands exponentially rapidly. Compute the horizon – the largest distance from which light can travel to an observer.

19

Astroparticle physics and inflation

In Chapter 18, we put forward a history of the universe. The picture is extremely simple. Its inputs are Einstein's equations and the assumptions of homogeneity and isotropy. We also used our knowledge of laws of atomic, nuclear and particle physics. We saw a number of striking confirmations of this basic picture, but there are many puzzles.

(1) The most fundamental problem is that we don't know the laws of physics relevant to temperatures greater than about 100 GeV. If there is only a single Higgs doublet at the weak scale, it is possible that we can extend this picture back to far earlier times. If there is, say, supersymmetry or large extra dimensions, the story could change drastically. Even if things are simple at the weak scale, we will not be able to extend the picture all the way back to $t = 0$. We have already seen that the classical gravity analysis breaks down.

(2) There are a number of features of the *present* picture we cannot account for within the Standard Model. Specifically, what is the dark matter? There is no candidate among the particles of the Standard Model. Is it some new kind of particle? As we will see, there are plausible candidates from the theoretical structures we have proposed, and they are the subject of intense experimental searches.

(3) The dark energy is very mysterious. Assuming it is a cosmological constant, it can be thought of as the vacuum energy of the underlying microphysical theory. As a number, it is totally bizarre. Its natural value should be set by the largest relevant scale, perhaps the Planck or unification scale, or the scale of supersymmetry breaking. Other proposals have been put forward to model the dark energy. One which has been extensively investigated is known as quintessence, the possibility that the energy is that of a slowly varying scalar field. Such models typically do not predict $w \neq -1$, and many are already ruled out by observations. But it should be stressed that these models are, if anything, less plausible than the possibility of a cosmological constant. First, one needs to explain why the underlying microphysical theory produces essentially zero cosmological constant, and a potential whose curvature is smaller than the present value of H. Then one needs to understand why the slowly varying field produces today the bizarre observed energy

density, without disturbing the successes of the cosmological picture for earlier times. It is probably fair to say that no convincing explanation of either aspect of the problem has been forthcoming.

(4) The value of the present baryon to photon ratio is puzzling:

$$\frac{n_{\rm B}}{n_\gamma} = \left(6.1\,^{+0.3}_{-0.2}\right) \times 10^{-10}. \tag{19.1}$$

As we will see, the question can alternately be phrased: why is this so small – or why is it so large? If the universe was always in thermal equilibrium, this number is a constant. So at very early times, there was a very tiny excess of particles over antiparticles. One might imagine that this is simply an initial condition, but, as A. Sakharov first pointed out, this is a number one might hope to explain through cosmology combined with microphysical theory. As we will discuss in detail later, it is necessary that the underlying microphysics violate baryon number and CP, and that there be a significant departure from thermal equilibrium. The Standard Model, as we have seen, violates both, and can generate a baryon number, but, as we will see, it is far too small. So *modifications of the known physical laws are required to account for the observed density of baryons.*

(5) Homogeneity, flatness and topological objects, such as monopoles, pose puzzles which suggest a phenomenon known as *inflation*. Consider, first, homogeneity. This certainly made the equations simple to solve, but it is puzzling. If we look at the cosmic microwave background, the temperature in different directions in the sky is equal to about a part in 10^5. But as we look out at distances 14 billion light years away, points separated by a tiny fraction of a degree were separated, at 100 000 years after the big bang, by an enormous distance compared to the horizon at that time. The problem is that the horizon decreases in size, as we look back, as $\sqrt{t}$. So points separated by a degree were, at that time, separated by about 10^7 light years. But signals could not travel more than 10^5 light years by this time. So if these points had not been in causal contact by recombination, why should they have identical temperatures? For nucleosynthesis, which occurs much earlier, the question is even more dramatic.

(6) Flatness ($\Omega_{\rm tot} = 1$) may not seem puzzling at first, but let's consider, again, the structure of the FRW metric. We have seen that the Friedmann equation can be recast as:

$$\frac{8\pi G\rho}{H^2} = \Omega - 1. \tag{19.2}$$

Suppose, for example, that $\Omega = 0.999$ today. Then, at recombination, the left-hand side of this equation was more than 8 orders of magnitude smaller. So the energy density was equal to the critical density with extraordinary precision. This apparent fine tuning gets more and more extreme as we look further back in time.

(7) Monopoles: we have seen that simple grand unified theories predict the existence of magnetic monopoles. Their masses are typically of order the grand unification scale. So unless their density is many orders of magnitude (perhaps 14!) smaller than the density of baryons, their total energy density will be far greater than the observed energy density of the universe. Astrophysical limits turn out to be even smaller; passing through the

galaxy, monopoles would deplete the magnetic field. This sets a limit, known as the Parker bound, on the monopole flux in the galaxy:

$$\mathcal{F} < 10^{-16} \left(\frac{M_{\rm mon}}{10^{17}\,{\rm GeV}} \right) {\rm cm}^{-2}\ {\rm s}^{-1}\ {\rm sr}^{-1}. \tag{19.3}$$

On the other hand, we might expect, in a grand unified theory, quite extensive monopole production. We have seen that monopoles are topological objects. If there is a phase transition between a phase of broken and unbroken $SU(5)$, we would expect twists in the fields on scales of order the Hubble radius at this time, and a density of monopoles of order one per horizon volume. If the transition occurs at $T_0 = 10^{16}$ GeV, the Hubble radius is of order $T^2/M_{\rm p}$, so the density, in units of the photon density, T^3, is of order

$$\frac{n_{\rm mon}}{n_\gamma} = \frac{T^3}{M_{\rm p}^3} \tag{19.4}$$

and can be *larger* than the baryon density.

In the following sections, we discuss these issues. We will study a possible solution to the homogeneity, flatness and monopole problems: inflation, the hypothesis that the universe underwent a period of extremely rapid expansion. We will see that there is some evidence that this phenomenon really occurred. Certainly there is nothing within the Standard Model itself which can give rise to inflation, so this points to the presence of some new phenomena, perhaps fields, perhaps more complicated entities, which are crucial to understanding the universe we see around us. We will describe some simple models of inflation, especially slow roll inflation, chaotic and hybrid inflation, and some of their successes and the puzzles which they raise. We will discuss inflationary theory's biggest success: quantum mechanical fluctuations during inflation give rise to the perturbations which grow to become the structure we see around us in the universe. This introduction is not comprehensive, but should give the reader some tools to approach the vast literature which exists on this subject.

We next turn to the problem of the dark matter. We focus on two candidates: the lightest supersymmetric particle of the MSSM, and the axion. We explain how these particles might rather naturally be produced with the observed energy density, and discuss briefly the prospects for their direct detection. Then we turn to baryogenesis. We explain why the Standard Model has all of the ingredients to produce an excess of baryons over anti-baryons, but, given the value of the Higgs mass, this baryon number cannot be nearly as large as is observed. We then turn to baryon production in some of our proposals for physics beyond the Standard Model, focussing on three possibilities: heavy particle decay in grand unified theories, leptogenesis, and coherent production by scalar fields.

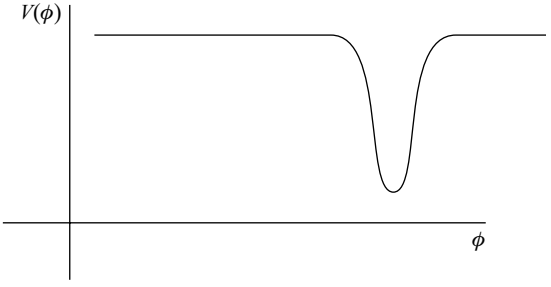

Fig. 19.1. A typical inflationary potential has a region in which $V(\phi)$ varies slowly and then settles into a minimum.

19.1 Inflation

The underlying idea behind inflation is that the universe behaved for a time as if (or nearly as if) the energy density was dominated by a positive cosmological constant, Λ. During this era, the Friedmann equation is that for de Sitter space,

$$H_{\rm i}^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\Lambda, \tag{19.5}$$

with solution

$$a(t) = e^{H_{\rm i} t}. \tag{19.6}$$

If this situation holds for a time such that, say, $\Delta t\, H_{\rm i} = 60$, then the universe expands by an enormous factor. Suppose, for example, $\Lambda = 10^{16}$ GeV; correspondingly $H_{\rm i} \approx 10^{14}$ GeV. Then a patch of size H^{-1} grows to be almost a centimeter in size. If, at the end of this period of inflation, the temperature of the universe is 10^{16} GeV, this patch would have grown, by the present time, by a factor of 10^{29}. This is about the size of our present horizon!

One possibility for how this might come about is called "slow roll inflation." Here one has a scalar field, ϕ, with potential $V(\phi)$. $V(\phi)$, for some range of ϕ, is slowly varying (Fig. 19.1). What we called $H_{\rm i}$ is determined by the average value of the potential in the plateau region, V_0. If we assume that we have a patch, initially, of size a bit larger than $H_{\rm i}^{-1}$, then we can write an equation of motion for the zero-momentum mode of the field, ϕ, in this region:

$$g^{\mu\nu} D_\mu \partial_\nu \phi + V'(\phi) = 0. \tag{19.7}$$

Because of our assumption of homogeneity (and isotropy), we can take the metric to have the Robertson–Walker form:

$$\ddot{\phi} + 3h\dot{\phi} + V'(\phi) = 0. \tag{19.8}$$

We assume that the field is moving slowly, so that we can neglect the $\ddot{\phi}$ term. Shortly, we will check whether this assumption is self-consistent. In this limit, the equation of motion is first order:

$$\dot{\phi} = -\frac{V'}{3H}. \tag{19.9}$$

We can integrate this equation to get Δt, the time it takes the field to traverse the plateau of the potential:

$$\Delta t = -\int d\phi \frac{3H(\phi)M_{\rm p}^2}{V'(\phi)}. \tag{19.10}$$

Assuming the validity of the slow roll approximation, the requirement for obtaining adequate inflation is:

$$N = \Delta t H > 60. \tag{19.11}$$

Now we can determine the conditions for the validity of the slow roll approximation. We simply want to check, from our solution, that $\ddot{\phi} \ll 3H\dot{\phi}, V'(\phi)$. Differentiating the solution, leads to the conditions

$$\epsilon = \frac{1}{2}M_{\rm p}^2\left(\frac{V'}{V}\right)^2 \ll 1 \tag{19.12}$$

and

$$\eta = M_{\rm p}^2\frac{V''}{V},\ |\eta| \ll 1. \tag{19.13}$$

How does inflation end? Near the minimum of the potential, we can approximate the potential as quadratic. So we might try to study an equation of the form:

$$\ddot{\phi} + 3H\phi + m^2\phi = 0. \tag{19.14}$$

Were it not for the expansion, this equation would have a solution

$$\phi = \phi_0 \cos(mt). \tag{19.15}$$

In a quantum mechanical language, this would describe a coherent state of particles, with energy density

$$\rho = \frac{1}{2}m^2\phi_0^2. \tag{19.16}$$

These particles have zero momentum; the pressure, $T_{ij} = p\delta_{ij} = 0$. So, if this field dominates the energy density of the universe, we know that

$$a \sim t^{2/3}; H = \frac{2}{3}t. \tag{19.17}$$

In our toy model, we might imagine $m \sim 10^{16}$ GeV $\gg H$, so we could solve the equation by assuming:

$$\phi(t) = f(t)\cos(mt) \tag{19.18}$$

for a slowly varying function, f. Plugging in, one finds

$$f(t) = \frac{1}{t}. \tag{19.19}$$

Note that this means

$$\rho = \rho_0 \left(\frac{t}{t_0}\right)^2 = \rho_0 \left(\frac{a}{a_0}\right)^3. \tag{19.20}$$

To summarize, we are describing a system which behaves like pressureless dust – zero-momentum particles – which is diluted by the expansion of the universe.

This description also gives us a clue as to the fate of the field, ϕ. Supposing that the ϕ particles have a finite width, Γ, they will decay in time $1/\Gamma$. We can include this in our equation of motion, writing:

$$\ddot{\phi} + (3H + \Gamma)\dot{\phi} + V'(\phi) = 0. \tag{19.21}$$

When the particles decay, if their decay products include, for example, ordinary quarks, leptons and gauge fields, their interactions will bring them quickly to equilibrium. We can be at least somewhat quantitative about this. The condensate disappears at a time set by $H \approx \Gamma$. If the universe quickly comes to equilibrium, the temperature must satisfy:

$$\frac{\pi^2}{30} g T^4 = H^2 \frac{3}{8\pi G}. \tag{19.22}$$

At this temperature, we can estimate the rate of interaction. Since the typical particle energy will be of order T, the cross sections will be of order

$$\sigma = \frac{\alpha_i^2}{T^2} \tag{19.23}$$

We can multiply this by the density, $n = (\pi^2/30) g^* T^3$, to obtain a reaction rate. For inflation with the scales we are discussing, this is enormous compared to H. The details by which equilibrium is established have been studied with some care. We can imagine that when a ϕ first decays, it produces two very-high-energy particles. These will have rather small cross sections for scattering with other high-energy decay products, but these interactions degrade the energy, and the cross sections for subsequent interactions – and for interactions with previously produced particles – are larger. More careful study leads to a behavior with time where the temperature

rises to a maximum and then falls. This maximum temperature is:

$$T_{\max} \approx 0.8 g_*^{-1/4} m^{1/2} (\Gamma M_{\rm p})^{1/4}, \tag{19.24}$$

where M is the mass of the inflaton.

19.1.1 Fluctuations: the formation of structure

One of the most exciting features of inflation is that it predicts that the universe is not exactly homogeneous and isotropic. We can't do justice to this subject in this short section, but can at least give the flavor of the analysis and collect the crucial formulae. In order to have inflation, we need that the metric and fields are reasonably uniform over a region of size $H_{\rm i}^{-1}$. But because of quantum fluctuations, the fields, and in particular the scalar field, ϕ, cannot be completely uniform. We can estimate the size of these quantum fluctuations without great difficulty. In order that inflation occur at all, we need $m_\phi \ll H$. So we will treat ϕ as a massless free field in de Sitter space. For such a field, we can estimate the size of quantum fluctuations. As in flat space, we can expand the field, ϕ, in Fourier modes:

$$\phi(\vec{x}, t) = \int \frac{d^3k}{(2\pi)^3} (e^{i\vec{k}\cdot\vec{x}} h(\vec{k}, t) + \text{c.c.}). \tag{19.25}$$

The expansion coefficients, h, obey the equation $D_\mu \partial_\nu h(\vec{k}, t) e^{i\vec{k}\cdot\vec{x}} = 0$, yielding, in the FRW background:

$$\ddot{h} + 3H\dot{h} + \frac{k^2}{a^2} h = 0. \tag{19.26}$$

Here k/a is the red-shifted momentum. In the case of de Sitter space, a grows exponentially rapidly. As soon as $k/a \sim H$, the system becomes overdamped, and the value of h is essentially frozen. We will see this in a moment when we write an explicit solution of the equation.

It is convenient to change our choice of time variable. Rather than take the FRW form for the metric, we take a metric more symmetric between space and time:

$$ds^2 = a^2(t)(-d\eta^2 + d\vec{x}^2). \tag{19.27}$$

Here, in terms of our original variables,

$$\frac{d\eta}{dt} = \pm \frac{1}{a}. \tag{19.28}$$

So, choosing the + sign,

$$\eta = \int \frac{da}{(\dot{a}/a)a^2} = \int \frac{da}{Ha^2} \tag{19.29}$$

and

$$\eta = H^{-1}/a. \tag{19.30}$$

In these coordinates, the equations of motion for $h(k, \eta)$ read:

$$\delta\ddot{\phi} + 2aH\delta\dot{\phi} + k^2\delta\phi = 0. \tag{19.31}$$

This equation is straightforward to solve. The solution can be written in terms of Bessel functions, but more transparently as:

$$\delta\phi_k = \frac{e^{-ik\eta}}{\sqrt{ik}}\left[1 - \frac{i}{k\eta}\right]. \tag{19.32}$$

Note that for large times, $\eta \to 0$.

Further analysis is required to convert this expression into a fluctuation spectrum. The result is that the fluctuations in the energy density are roughly scale-invariant, and

$$\frac{\delta\rho}{\rho} \approx \left(\frac{H^2}{5\pi\dot{\phi}}\right). \tag{19.33}$$

Using the slow roll equation,

$$3H\dot{\phi} = V', \tag{19.34}$$

gives

$$\frac{\delta\rho}{\rho} = \frac{3H^*}{5\pi V'} = \frac{3V^{3/2}}{5\pi V' M_{\rm p}^3}. \tag{19.35}$$

Much more detailed discussion of these formulas can be found in the suggested reading. These fluctuations quickly pass out of the horizon during inflation. While outside of the horizon, they are frozen. Subsequently, however, they reenter the horizon and begin to grow. Measurements of the CMBR indicate that

$$\frac{\delta\rho}{\rho} \sim 10^{-5} \tag{19.36}$$

on horizon scales. Fluctuations which were within the horizon at the time of matter-radiation equality have grown linearly with time since that time. At about 1 billion years after the big bang, they became non-linear, and this appears to account adequately for the observed structure in the universe. Precise studies of the CMBR, the formation of structure, and Type Ia supernovas, as well as of the missing mass in structures on a wide range of scales, has allowed the determination of the composition of the universe.

But while the overall scenario is compelling, and has significant observational support, we lack a persuasive microphysical understanding of these phenomena.

This is undoubtedly one of the great challenges of theoretical physics. In the next section, we describe two classes of models, each of which can be motivated by considerations of supersymmetry and of string theory, for inflation.

19.1.2 Models of inflation

There are almost as many models of inflation as there are physicists who have thought about the problem, and we cannot possibly sample all of them. In this section we survey a few. So far, we have discussed a model of a single scalar field. The properties of the potential for this field must be finely tuned in order to obtain enough e-foldings of inflation and small enough $\delta\rho/\rho$. First, it is known from observations that the Hubble constant during inflation cannot be much larger than 10^{16} GeV. This means that the scalar mass cannot be comparable to the Planck mass, so we face the usual problem of light scalars. Second, the existence of a light scalar by itself does not insure adequate e-foldings and fluctuation spectrum without further fine tuning. The first models of slow roll inflation, based on what is known as the "Coleman–Weinberg" potential, were severely fine-tuned in both senses. We can illustrate the problems with another class of models, known as "chaotic inflation." The idea is to consider a single scalar field, with a simple potential, such as $\lambda\phi^4$. This model can produce inflation if λ is very small, and the typical values of ϕ are very large. The validity of the slow roll picture requires that $\phi \gg M_{\rm p}$. For definiteness, suppose $\phi \approx 10 M_{\rm p}$. Then

$$\frac{\delta\rho}{\rho} \approx \lambda^{1/2}\phi^2 \tag{19.37}$$

so that $\lambda \sim 10^{-16}$. With this choice, one obtains, from Eq. (19.10), over 1000 e-foldings during inflation.

This model is very simple, but the extremely small value of the coupling is troubling, especially given the large value of ϕ. Not only is this small value of the coupling puzzling, but the ϕ^n couplings, $n > 4$, expressed in Planck units, must be even smaller. Despite these concerns, this model has proven useful for considering many aspects of inflation, and it has been argued that some of its features may characterize a larger class of models.

Given that supersymmetry naturally produces light scalars, supersymmetry would seem a natural context in which to construct models of inflation. We have mentioned that in supersymmetric field theories and in string theory, one often encounters moduli, scalar fields whose potentials vanish in the some limit. Banks has suggested that for such fields, a potential of the form

$$V = \mu^4 f(\phi/M_{\rm p}) \tag{19.38}$$

will often arise. Here μ is an energy scale determined by some dynamical phenomenon such as the scale of supersymmetry breaking. For such a potential, assuming that typical field values are of order $M_{\rm p}$, we have from Eq. (19.35),

$$\frac{\delta\rho}{\rho} \approx \mu^2/M_{\rm p}^2. \tag{19.39}$$

From this, we have $\mu \approx 10^{15.5}$ GeV. The number of e-foldings is generically of order one; the potential must be tuned to the level of 1%, for example, if one is to obtain sufficient inflation. Still, this may seem less troubling than having many couplings less that 10^{-6}. Note that μ, the energy scale in a supersymmetric model of this kind with a single field is far larger than we have considered for low-energy supersymmetry breaking. Banks proposes that at the minimum of the potential, supersymmetry is unbroken, with vanishing $\langle W\rangle$ as a result of an R-symmetry.

Another class of models of some interest are known as hybrid models. These involve two fields. They are particularly interesting in the context of supersymmetry, where the scale of inflation can now be quite low, even comparable to the TeV scale. Here we consider an example from a class of models dubbed by Guth and Randall "supernatural," since the presence of light scalars is again natural. The model contains two chiral fields, χ and ϕ, responsible for inflation. Both χ and ϕ, initially, sit far from their minima, in such a way that the field χ is very massive, but ϕ is light (more precisely, the curvature of the χ as potential is very large, but that of the ϕ potential is of order the basic scale of the model). During this period, χ is essentially frozen, while ϕ moves slowly towards a local minimum. The energy is dominated by the potential for χ. As ϕ approaches its minimum, the χ field becomes lighter, and it eventually begins to roll towards *its* minimum. At this time, inflation ends. The universe reheats as χ oscillates about the minimum of the potential.

In more detail, we can take the superpotential to be

$$W = \frac{\lambda\chi\phi^3}{M} \tag{19.40}$$

and a soft breaking potential

$$V_{\rm soft}(\phi, \chi) = m_{3/2}^2 g(\phi) + \mathcal{N} m_{3/2}^2 f(\chi). \tag{19.41}$$

We assume that the global minimum of g is at the origin, but that f has a local minimum near $\chi = 0$, its global minimum being located a distance of order $M_{\rm p}$ away.

The mass of the χ field receives a contribution from the superpotential term, which dominates until $\phi^2 \approx \mathcal{N} m_{3/2} M_{\rm p}$. When ϕ reaches this point, χ starts to move towards its true minimum. Unless the potential is further tuned, χ will reach its

minimum in a few Hubble times, ending inflation; χ oscillates about its minimum, leading to a period of matter domination, until it decays, reheating the universe.

The number of e-foldings, using Eq. (19.10), from large ϕ until ϕ reaches this critical value is of order $\mathcal{N}$. The fluctuation amplitude is of order

$$\frac{\delta\rho}{\rho} \sim \mathcal{N}\left(\frac{m_{3/2}}{M_{\rm p}}\right)^{1/2} \approx \mathcal{N} \times 10^{-7.5}. \tag{19.42}$$

So $\mathcal{N} \sim 60$ gives a suitable number of e-foldings and a reasonable fluctuation spectrum. If the mass of χ is of order $\mathcal{N}^{1/2}$ TeV, the χ width is of order:

$$\Gamma_\chi = \frac{m_\chi^3}{M_{\rm p}^2} \propto \mathcal{N}^{3/2}. \tag{19.43}$$

The corresponding reheating temperature, if $\mathcal{N} \sim 60$ is of order $T_{\rm r} = 10$ MeV. This is above nucleosynthesis, but just barely. It is compatible with everything we know. In such a picture, the thermal history of the universe began just before nucleosynthesis; there was not a phase transition in QCD or in the weak interactions, for example.

19.1.3 Constraints on reheating: the gravitino problem

Our discussion of supersymmetric models of inflation suggests a broad range of possible reheating temperatures, $T_{\rm r}$. But there is a reason, at least in the context of supersymmetric theories, to think that there may be an upper bound on the reheating temperature. This is the problem of producing too many gravitinos. The gravitino lifetime is quite long,

$$\Gamma_{3/2} \approx \frac{m_{3/2}^3}{M_{\rm p}^2}; \tag{19.44}$$

gravitinos might even be stable. As a minimal requirement, we need to suppose that the gravitinos did not dominate the energy density at the time of nucleosynthesis. If they did, again, they spoil the successful predictions of nucleosynthesis, but worse, as for moduli, their decay products destroy the light elements. Even though gravitinos are very weakly interacting, there is a danger that they will be overproduced during the period of reheating that follows inflation. A natural estimate for their production rate per unit volume is obtained by assuming that they are produced in two-body scattering, by light particles with densities of order T^3, and that their cross sections behave as $1/M_{\rm p}^2$:

$$n^2\langle\sigma v\rangle \approx T^6\langle\sigma v\rangle \approx \frac{T^6}{M_{\rm p}^2}. \tag{19.45}$$

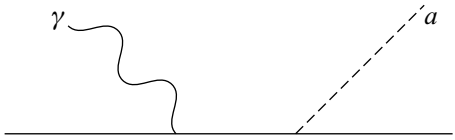

Fig. 19.2. In a Bremsstrahlung-like process, a lepton or nucleon can emit an axion when struck by a photon.

Integrating this over a Hubble time, $M_{\rm p}/T^2$, and dividing by the photon density, of order T^3 gives a rough estimate:

$$\frac{n_{3/2}}{s} \sim \frac{T}{M_{\rm p}}. \tag{19.46}$$

Assuming 1 TeV for the gravitino mass, the requirement that gravitinos not dominate before nucleosynthesis gives $T < 10^{12}$ GeV. But this is too crude. Considering destruction of deuterium and lithium gives $T < 10^9$ GeV and possibly much smaller. This is a strong constraint on the nature of reheating after inflation. In the models suggested by Banks, for example, the reheating temperature can easily be larger. This is not a problem for the low-scale, hybrid models we discussed in the previous section.

19.2 The axion as dark matter

Within the set of ideas we have discussed for physics beyond the Standard Model, there are two promising candidates for dark matter. One is the axion, which we discussed in Chapter 5 as a possible solution to the strong CP problem. A second is the lightest supersymmetric particle in models with an unbroken R-parity. We first discuss the axion, mainly because the theory is particularly simple. To begin, we need to consider the astrophysics of the axion a bit further. There is a lower bound on the axion decay constant, or equivalently an upper bound on the axion mass, arising from processes in stars. Axions can be produced by collisions deep within the star. Then, because of their small cross section, most axions will escape, carrying off energy. This has the potential to disrupt the star. We can set a limit by requiring that the flux of energy from the stars not be more than a modest fraction of the total energy flux.

To estimate these effects, we can first ask what sorts of processes might be problematic. A pair of photons can collide and produce an axion (using the $a F \tilde{F}$ coupling of the axion to the photon). Axions can be produced from nuclei or electrons in Compton-like and Bremsstrahlung processes (Fig. 19.2). The typical energies will be of order T.

For the Compton-like process of Fig. 19.2, the cross section is of order:

$$\sigma_{\rm a} \approx \frac{\alpha}{f_{\rm a}^2}. \tag{19.47}$$

The total rate per unit time for a given electron to scatter off a photon in this way will be proportional to the photon density, which we will simply approximate as T^3. To obtain the total emission per unit volume, we need to multiply, as well, by the electron density in the star. In the Sun, this number is of order the total number of protons or electrons, 1.16×10^{57}, divided by the cube of the solar radius (in particle physics units, 3.5×10^{25} GeV^{-1}). This corresponds to

$$n_{\rm e} \approx 3 \times 10^{-16}\ ({\rm GeV})^3 \text{ electrons}. \tag{19.48}$$

Rather than calculate the absolute rate, let's compare with the rate for neutrino production. We would expect that if axions carry off far more energy than neutrinos, this is problematic. For neutrino production we might take $n_{\rm e}^2$, and multiply by a typical weak cross section,

$$\sigma_\nu = G_{\rm F}^2 E^2. \tag{19.49}$$

Finally, we take the temperature in the core of the star to be of order 1 MeV. Taking $f_{\rm a} = 10^9$, gives, for the axion production rate:

$$R_{\rm axion} = 10^{-47}\ {\rm GeV}^{-4} \tag{19.50}$$

while

$$R_\nu = 10^{-47}\ {\rm GeV}^{-4}. \tag{19.51}$$

Clearly this analysis is crude; much more care is required in enumerating different processes and evaluating their cross sections, and integrating over particle momentum distributions. But this rough calculation indicates that 10^9 GeV is a plausible lower limit on the axion decay constant.

So we have a lower bound on the axion decay constant. An upper bound arises from cosmology. Suppose that the Peccei–Quinn symmetry breaks before inflation. Then throughout what will be the observable universe, the axion is essentially constant. But at early times, the axion potential is negligible. To be more precise, consider the equation of motion of the axion field:

$$\ddot{a} + 3H\dot{a} + V'(a) = 0. \tag{19.52}$$

At very early times, $H \gg m$, and the system is overdamped. The axion simply doesn't move. If the universe is very hot, the axion mass is actually much smaller than its current value. This is explained in Appendix C, but is easy to understand. At very high temperatures, the leading contribution in QCD to the axion potential

comes from instantons. Instanton corrections are suppressed by $e^{\frac{-8\pi^2}{g^2(T)}} = (\Lambda/T)^{b_0}$. They are also suppressed by powers of the quark masses. In other words, they behave as:

$$V(a) = \prod_f m_{\rm f} \Lambda^{b_0} T^{-b_0+n_{\rm f}-4} \cos(\theta) \tag{19.53}$$

where $\theta = a/f_{\rm a}$, and $n_{\rm f}$ is the number of flavors with mass $\ll T$. This goes very rapidly to zero at temperatures above the QCD scale.

So the value of the axion field – the θ-angle – at early times, is most likely simply a random variable. Let's consider, then, the subsequent evolution of the system. The equation of motion for such a scalar field in a FRW background is by now quite familiar:

$$\ddot{a} + H\dot{a} + V'(a) = 0. \tag{19.54}$$

$V(a)$ also depends on $T(t)$, which complicates slightly the solution, so let's first just solve the problem with the zero-temperature axion potential. In this case, the axion will start to oscillate when $H \sim m_{\rm a}$. After this, the axions dilute like matter, i.e. as $1/a^3$. The energy in radiation, on the other hand, dilutes like $a^4 \propto T^{-4}$. Assuming radiation domination when the axion starts to oscillate, we can determine the temperature at that time. Using our standard formula for the energy density,

$$\rho = \frac{\pi^2}{30} g^* T^4 \tag{19.55}$$

we have, just above the QCD phase transition, $g^* \approx 48$ (with the gluons, three quark flavors, three light neutrinos and the photon). Just below, we don't have the quarks or gluons, but we should include the pions, so $g^* \approx 30$. Taking the larger value

$$T_{\rm a} = 10^2 \text{ GeV} \left(\frac{10^{11}\text{GeV}}{f_{\rm a}}\right)^{1/2}. \tag{19.56}$$

At this time, the fraction of the energy density in axions is approximately

$$\frac{\rho_{\rm axion}}{\rho} = \frac{\frac{1}{2} f_{\rm a}^2 m_{\rm a}^2}{\rho} \approx \frac{1}{6} \frac{f_{\rm a}^2}{M_{\rm p}^2}. \tag{19.57}$$

So if $f_{\rm a} = 10^{11}$ GeV, axions come to dominate the energy density quite late, at $T \approx 10^{-3}$ eV. The temperature of axion domination scales with $f_{\rm a}$, so a 10^{16} GeV axion would dominate the energy density at 100 eV, which would be problematic.

However, the axion potential, as we have seen, is highly suppressed at temperatures above a few hundred MeV. So oscillation, in fact, sets in much later. We can make another crude estimate by simply supposing that the axion potential turns on at $T = 100$ TeV. In this case, the axion fraction is large, of order $1/g^*$. So if

the axion density is to be compatible with the observed dark matter density, for any value of f_a, we need to allow for the detailed temperature dependence of the axion mass. Using our formula for the axion potential as a function of temperature we can ask when the associated mass becomes of order the Hubble constant. After that time, axion oscillations are more rapid than the Hubble expansion, so we might expect that the axion density will damp, subsequently, like matter. Let's take, specifically, $f_a = 10^{11}$ GeV. For the axion mass, we can take:

$$m_a(T) \approx 0.1 m_a(T=0) \left(\frac{\Lambda_{QCD}}{T} \right)^{3.7}. \tag{19.58}$$

The axion then starts to oscillate when $T \approx 1.5$ GeV. At this time, it represents about 10^{-9} of the energy density. One needs to do a bit more work to show that in the subsequent evolution, the energy in axions relative to the energy in radiation falls roughly as $1/T$, but for this decay constant, the axion and radiation energies become equal at roughly 1 eV. If the decay constant is significantly higher than 10^{11} GeV, then the axions start to oscillate too late, and dominate the energy density too early. If the decay constant is significantly smaller, then the axions cannot constitute the presently observed dark matter.

So it is remarkable that there is a rather narrow range of axion decay constants consistent with observation. On the other hand, some of the assumptions we have made in this section are open to question. In the case of hybrid inflation we have seen that the universe might never have been much hotter than 10 MeV. In this case, the upper limit on the axion decay constant, as we will discuss further later, can be much weaker.

19.3 The LSP as the dark matter

A stable particle is not necessarily a good dark matter candidate. But we can make a crude calculation which indicates that the LSP density is in a suitable range to be the dark matter. Consider particles, X, with mass of order 100 GeV interacting with weak interaction strength. Their annihilation and production cross sections go as $G_F^2 E^2$. So, in the early universe, the corresponding interaction rate is of order

$$\Gamma \approx \rho_X G_F^2 E^2 \approx \rho_X G_F^2 T^2. \tag{19.59}$$

These interactions will drop out of equilibrium when the mass of the particle X is small compared to the temperature, so that there is a large Boltzmann suppression of their production. This will occur when this rate is of order the expansion rate, or

$$T^3 e^{-M_X/T} \langle v\sigma \rangle \sim \frac{T^2}{M_p}. \tag{19.60}$$

Since the exponent is very small, once $T \sim 10 M_X$, we can get a rough estimate of the density by saying that:

$$e^{-M_X/T} T^3 \sim G_{\rm F}^{-2}/M_{\rm p}. \tag{19.61}$$

The ratio, n_X/s, then, is of order

$$n_X/s \approx G_{\rm F}^{-2}/(M_{\rm p} T^3). \tag{19.62}$$

Assuming that $M_X \sim 100$ GeV, and $T \sim 10\,\text{GeV}$, this gives about 10^{-9} for the right-hand side. Since the energy density in radiation damps like T^{-4}, while that for matter damps like T^{-3}, this gives matter-radiation equality at temperatures of order an electronvolt, as in the standard big bang cosmology.

Needless to say, this calculation is quite crude. Extensive, far more sophisticated, calculations have been done to find the regions of parameter space in different supersymmetric models which are compatible with the observed dark matter density. The basic starting point for these analyses is the Boltzmann equation. If the basic process is of the form $1 + 2 \leftrightarrow 3 + 4$, then:

$$a^{-3} \frac{d}{dt}(n_1 a^3) = \int \frac{d^3 p_1}{(2\pi)^3 2E_1} \int \frac{d^3 p_2}{(2\pi)^3 2E_2} \int \frac{d^3 p_3}{(2\pi)^3 2E_3} \int \frac{d^3 p_4}{(2\pi)^3 2E_4}$$

$$(f_3 f_4[1 \pm f_1][1 \pm f_2] - f_1 f_2[1 \pm f_3][1 \pm f_4])(2\pi)^4 \delta^4(p_1 + p_2 - p_3 - p_4)|\mathcal{M}|^2. \tag{19.63}$$

The functions $f_1, \ldots, f_4$ are the distribution functions for the different species. These equations can be simplified in the high-temperature limit using Boltzmann statistics:

$$f(E) \to e^{\mu/T} e^{-E/T}. \tag{19.64}$$

Interactions are still fast enough at this time to maintain equilibrium of the X momentum distributions (kinetic equilibrium) but not that of X number. So it is the limiting value of the X chemical potential, μ_X, which we seek. In this limit, we have:

$$f_3 f_4[1 \pm f_1][1 \pm f_2] - f_1 f_2[1 \pm f_3][1 \pm f_4]$$

$$\to e^{-(E_1+E_2)/T}\left(e^{(\mu_1+\mu_2)/T} - e^{(\mu_3+\mu_4)/T}\right). \tag{19.65}$$

Here we have used $E_1 + E_2 = E_3 + E_4$.

Things simplify further as all but the X particle (particle 1) are light, and nearly in equilibrium. Defining $n_i^{(0)}$ as the distributions in the absence of a chemical potential, and defining the thermally averaged cross section:

$$\langle \sigma v \rangle = \frac{1}{n_1^{(0)} n_2^{(0)}} \int \frac{d^3 p_1}{(2\pi)^3 2E_1} \int \frac{d^3 p_2}{(2\pi)^3 2E_2} \int \frac{d^3 p_3}{(2\pi)^3 2E_3} \int \frac{d^3 p_4}{(2\pi)^3 2E_4} |\mathcal{M}|^2, \tag{19.66}$$

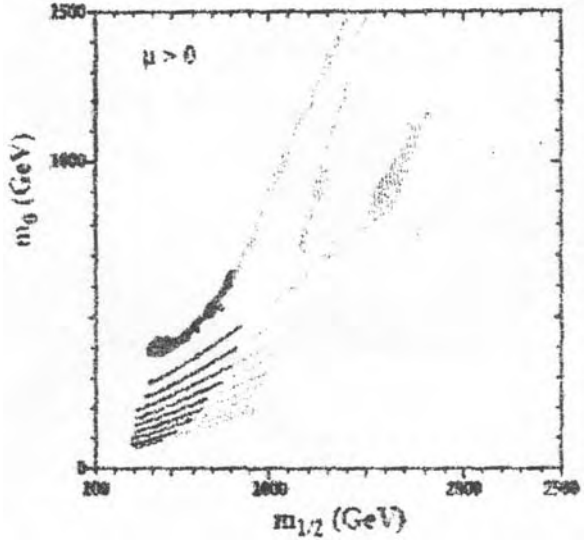

Fig. 19.3. Reprinted from J. R. Ellis *et al.*, Supersymmetric dark matter, *Phys. Lett. B*, **565**, 176 (2003), Figure 9. Copyright 2003, with permission from Elsevier.

we have

$$a^{-3}\frac{d(n_x a^3)}{dt} = n_X^{(0)} n_2 \langle \sigma v \rangle \left(1 - \frac{n_X}{n_X^{(0)}}\right). \tag{19.67}$$

Detailed solutions of these equations (often without some of these simplifications) reveal, as one would expect, a range of parameters in the MSSM compatible with the observed dark matter density (Fig 19.3).

So while it is disturbing that we need to impose additional symmetries in the MSSM in order to avoid proton decay, it is also exciting that this leads to a possible solution of one of the most critical problems of cosmology: the identity of the dark matter.

19.4 The moduli problem

We have seen that, in supersymmetric theories, there are frequently light moduli. In string models, we will find that such fields are ubiquitous. Such moduli, if they exist, pose a cosmological problem with some resemblance to the problems of axion cosmology.

In this section, we will formulate the problem as it arises in gravity-mediated supersymmetry breaking. The potential for the modulus, ϕ, would be expected to take the form:

$$V(\phi) = m_{3/2}^2 M_{\rm p}^2 f(\phi/M_{\rm p}). \tag{19.68}$$

By assumption, f has a minimum at some value, ϕ, of order $M_{\rm p}$. In the early universe, when the Hubble parameter is much greater than $m_{3/2}$, this potential is effectively quite small, and there is in general no obvious reason that the field should sit at its minimum. So, when $H \sim m_{3/2}$, the field is likely to lie at a distance $M_{\rm p}$, in field space from the minimum, and store an energy of order $m_{3/2}^2 M_{\rm p}^2$. Like the axion, after this time, assuming it is within the domain of attraction of the minimum, it will oscillate, behaving like pressureless dust. Almost immediately, given our assumptions about scales, it comes to dominate the energy density of the universe, and it continues to do so until it decays. The problem is that the decay occurs quite late, and the temperature after the decay is likely to be quite low.

We can estimate the temperature after ϕ decay, $T_{\rm r}$, by considering first the lifetime of the ϕ particle. We might expect this to be:

$$\Gamma_\phi = \frac{m_{3/2}^3}{M_{\rm p}^2} \tag{19.69}$$

assuming that the couplings of the ϕ field to other light fields are suppressed by a single power of M. Assuming that the decay products quickly thermalize, and noting that Γ_ϕ is the Hubble constant at the time of ϕ decay, gives

$$T_{\rm r}^4 \approx \frac{m_{3/2}^6}{M_2} \approx (10\ {\rm keV})^4. \tag{19.70}$$

Here we are assuming $m_{3/2} \approx 1$ TeV. This is a temperature well below nucleosynthesis temperature. So in such a picture, the universe is matter dominated during nucleosynthesis. But the situation is actually far worse: the decay products almost certainly destroy deuterium and the other light nuclei.

Two plausible resolutions for this puzzle have been put forward (apart from the obvious one that perhaps there is no supersymmetry or no moduli): the moduli might be significantly more massive than 1 TeV. Note that $T_{\rm r}$ scales like the moduli masses to the $3/2$ power, so if the moduli masses are of order hundreds of TeV, this temperature can be sufficiently high that nucleosynthesis occurs (again). There is, potentially, a serious problem with such a picture. During the ϕ decay, significant entropy is produced. For example, if the temperature of the universe when the moduli came to dominate was $\sqrt{m_{3/2} M_{\rm p}}$, when the moduli decay it is far smaller than $T_{\rm r}$. Correspondingly, the entropy is increased by a factor which can easily be 10^9–10^{12}. Any baryon number produced before these decays is diluted by a corresponding factor. One can hope that the baryons are produced in the decays of these moduli, but this requires one understand why such low-energy baryon-number violation doesn't cause difficulties for proton decay. In the next section, we will discuss possible mechanisms to produce the baryon asymmetry, and we will see

that there is one which is capable of producing a large enough asymmetry to survive moduli decays.

19.5 Baryogenesis

The baryon to photon ratio, n_B/n_γ, is quite small. At early times, when QCD was nearly a free theory, this slight excess would have been extremely unimportant. But for the structure of our present universe, it is terribly important. One might imagine that n_B/n_γ is simply an initial condition, but it would be more satisfying if we could have some microphysical understanding of this asymmetry between matter and antimatter. Andrei Sakharov, after the experimental discovery of CP violation, was the first to state precisely the conditions under which the laws of physics could lead to a prediction for the asymmetry.

(1) *The underlying laws must violate baryon number*. This one is obvious; if there is, for example, no net baryon number initially, and if baryon number is not conserved, the baryon number will always be zero.
(2) *The laws of nature must violate CP*. Otherwise, for every particle produced in interactions, an antiparticle will be produced as well.
(3) *The universe, in its history, must have experienced a departure from thermal equilibrium*. Otherwise, the CPT theorem insures that the numbers of baryons and anti-baryons at equilibrium are zero. This can be proven with various levels of rigor, but one way to understand this is to observe that CPT insures that the masses of the baryons and anti-baryons are identical, so at equilibrium their distributions should be the same.

Subsequently, there have been many proposals for how the asymmetry might arise. In the next sections, we will describe several. *Leptogenesis* relies on lepton-number violation, something we know is true of nature, but of whose underlying microphysics we are ignorant. *Baryogenesis through coherent scalar fields* (Affleck–Dine baryogenesis) also seems plausible. It is only operative if supersymmetry is unbroken to comparatively low energies, but it can operate quite late in the evolution of the universe and can be extremely efficient. This could be important in situations like moduli decay or hybrid inflation where the entropy of the universe is produced very late, after the baryon number.

19.5.1 Baryogenesis through heavy particle decays

One well-motivated framework in which to consider baryogenesis is grand unification. Here one can satisfy all of the requirements for baryogenesis. Baryon-number violation is one of the hallmarks of GUTs, and these models possess various sources of CP violation. As far as departure from equilibrium, the massive gauge bosons,

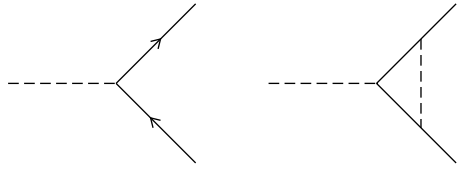

Fig. 19.4. Tree and loop diagrams whose interference can lead to an asymmetry in heavy particle decay.

X, provide good candidates for a mechanism. To understand in a bit more detail how the asymmetry can come about, note that CPT requires that the total decay rate of X is the same as that of its antiparticle $\bar{X}$. But it does not require equality of the decays to particular final states (partial widths). So starting with equal numbers of X and $\bar{X}$ particles, there can be a slight asymmetry between processes such as

$$X \to dL; X \to \bar{Q}\bar{u} \tag{19.71}$$

and

$$\bar{X} \to \bar{d}\bar{L}; \bar{X} \to Qu. \tag{19.72}$$

The tree graphs for these processes are necessarily equal; any CP-violating phase simply cancels out when we take the absolute square of the amplitude (see Fig. 19.4). This is not true in higher order, where additional phases associated with real intermediate states can appear. Actually computing the baryon asymmetry requires an analysis of the Boltzmann equations, of the kind he have encountered in our discussion of dark matter.

There are reasons to believe, however, that GUT baryogenesis is not the origin of the observed baryon asymmetry. Perhaps the most compelling of these has to do with inflation. Assuming that there was a period of inflation, any pre-existing baryon number was greatly diluted. So in order that one produces baryons through X boson decay, it is necessary that the reheating temperature after inflation be at least comparable to the X boson mass. But as we have explained, a reheating temperature greater than 10^9 GeV leads to cosmological difficulties, especially overproduction of gravitinos.

19.5.2 Electroweak baryogenesis

The Standard Model, for some range of parameters, can satisfy all of the conditions for baryogenesis. We have seen in our discussion of instantons that the Standard Model violates baryon number. This violation, we saw, is extremely small at low temperatures, so small that it is unlikely that a single baryon has decayed in the history of the universe this way. The rate is so small because baryon-number violation

is a tunneling process. If one could excite the system to high energies, one might expect that the rate would be enhanced. At high enough energies, the system might simply be above the barrier. One can find the configuration which corresponds to sitting on top of the barrier by looking for static, unstable solutions of the equations of motion. Such a solution is known. It is called a sphaleron (from Greek, meaning "ready to fall"). The barrier is quite high – from familiar scaling arguments, the sphaleron energy is of order $E_{\rm sp} = 1/\alpha M_W$. But this configuration is large compared to its energy; it has size of order M_W. As a result, it is difficult to produce in high-energy scattering. Two particles with enough energy to produce the sphaleron have momenta much higher than M_W. As a result, their overlap with the sphaleron configuration is exponentially suppressed.

At high temperatures one might expect that the sphaleron rate would be controlled by a Boltzmann factor, $e^{-E_{\rm sp}/T}$. So as the temperature increases, the rate would grow significantly. This turns out to be the case. In fact, the rate is even larger than one might expect from this estimate, because $E_{\rm sp}$ itself is a function of T. At very high temperatures, the rate has no exponential suppression at all, and behaves as:

$$\Gamma = (\alpha_w T)^4. \tag{19.73}$$

These phenomena are discussed in Appendix C.

If the Higgs mass is not too large, the Standard Model can produce a significant departure from equilibrium. As one raises the temperature, a simple calculation, described in Appendix C, shows that the Higgs mass increases (the mass-squared becomes less negative) with temperature. At very high temperature, the $SU(2) \times U(1)$ symmetry is restored. The phase transition between these two phases, for a sufficiently light Higgs, is first order. It proceeds by the formation of bubbles of the unbroken phase. The surfaces of these bubbles can be sites for baryon number production. These phenomena are also discussed in Appendix C. So the third of Sakharov's conditions can be satisfied.

Finally, we know that the Standard Model violates CP. We also know, however, that it is crucial that there are three generations, and that this CP violation vanishes if any of the quark masses are zero. As a result, even if the Higgs mass is small enough that the transition is strongly first order, any baryon number produced is suppressed by several powers of Yukawa couplings, and is far too small to account for the observed matter–antimatter asymmetry.

In the MSSM, the situation is somewhat better. There is a larger region of the parameter space in which the transition is first order, and as we have seen, there are many new sources of CP violation. As a result, there is, as of this writing, a small range of parameters where the observed asymmetry could be produced in this way.

19.5.3 Leptogenesis

There is compelling evidence that neutrinos have mass. The most economical explanation of these masses is that they arise from a seesaw, involving gauge singlet fermions, N_a. These couplings violate lepton number. So according to Sakharov's principles, we might hope to produce a lepton asymmetry in their decays. Because the electroweak interactions violate baryon and lepton number at high temperatures, the production of a lepton number leads to the production of baryon number.

In general, there may be several N_a fields, with couplings of the form:

$$\mathcal{L}_N = M_{ab} N_a N_b + h_{ai} H L_i N_a + \text{ c.c.} \tag{19.74}$$

In a model with three Ns, there are CP-violating phases in the Yukawa couplings of the Ns to the light Higgs. The heaviest of the right-handed neutrinos, say N_1, can decay to ℓ and a Higgs, or to $\bar{\ell}$ and a Higgs. At tree level, as in the case of GUT baryogenesis, the rates for production of leptons and anti-leptons are equal, even though there are CP-violating phases in the couplings. It is necessary, again, to look at quantum corrections, in which dynamical phases can appear in the amplitudes. At one loop, the decay amplitude for N has a discontinuity associated with the fact that the intermediate N_1 and N_2 can be on shell (similar to Fig. 19.4). So one obtains an asymmetry proportional to the imaginary parts of the Yukawa couplings of the Ns to the Higgs:

$$\epsilon = \frac{\Gamma(N_1 \to \ell H_2) - \Gamma(N_1 \to \bar{\ell} \bar{H}_2)}{\Gamma(N_1 \to \ell H_2) + \Gamma(N_1 \to \bar{\ell} \bar{H}_2)} = \frac{1}{8\pi} \frac{1}{h h^\dagger} \sum_{i=2,3} \text{Im}[(h_\nu h_\nu^\dagger)_{1i}]^2 f\left(\frac{M_i^2}{M_1^2}\right), \tag{19.75}$$

where f is a function that represents radiative corrections. For example, in the Standard Model $f = \sqrt{x}[(x-2)/(x-1) + (x+1)\ln(1+1/x)]$, while in the MSSM $f = \sqrt{x}[2/(x-1) + \ln(1+1/x)]$. Here we have allowed for the possibility of multiple Higgs fields, with H_2 coupling to the leptons. The rough order of magnitude here is readily understood by simply counting loops factors. It need not be terribly small.

Now, as the universe cools through temperatures of order of masses of the Ns, they drop out of equilibrium, and their decays can lead to an excess of neutrinos over antineutrinos. Detailed predictions can be obtained by integrating a suitable set of Boltzmann equations. But a rough estimate can be obtained by noting that the N_as drop out of equilibrium when their production rate becomes comparable to the expansion rate of the universe. If α represents a typical coupling, this occurs roughly when

$$\pi \alpha^2 T e^{-M_N/T} \approx \frac{T^2}{M_\text{p}}. \tag{19.76}$$

Assuming that in the polynomial terms, $T \sim M_N/10$, gives that the density at this time is of order

$$\frac{\rho_N}{\rho_{\rm tot}} \sim \frac{\pi T}{M_{\rm p}\alpha^2}. \tag{19.77}$$

Multiplying by ϵ, the average asymmetry in N decays, this estimate suggests a lepton number – and hence a baryon number – of order:

$$\frac{\rho_{\rm B}}{\rho_{\rm tot}} \approx \epsilon \frac{M_N}{10\pi\alpha^2 M_{\rm p}}. \tag{19.78}$$

We have seen that ϵ is suppressed by a loop factor and by Yukawa couplings. So this number can easily be compatible with observations, or even somewhat larger, depending on a variety of unknown parameters.

These decays, then, produce a net lepton number, but not baryon number (and hence a net $B - L$). The resulting lepton number will be further processed by sphaleron interactions, yielding a net lepton and baryon number (recall that sphaleron interactions preserve $B - L$, but violate B and L separately). One can determine the resulting asymmetry by an elementary thermodynamics exercise. One introduces chemical potentials for each neutrino, quark and charged lepton species. One then considers the various interactions between the species at equilibrium. For any allowed chemical reaction, the sum of the chemical potentials on each side of the reaction must be equal. For neutrinos, the relations come from:

(1) the sphaleron interactions themselves

$$\sum_i \left(3\mu_{q_i} + \mu_{\ell_i}\right) = 0; \tag{19.79}$$

(2) a similar relation for QCD sphalerons

$$\sum_i \left(2\mu_{q_i} - \mu_{u_i} - \mu_{d_i}\right) = 0; \tag{19.80}$$

(3) vanishing of the total hypercharge of the universe

$$\sum_i \left(\mu_{q_i} - 2\mu_{\bar{u}_i} + \mu_{\bar{d}_i} - \mu_{\ell_i} + \mu_{\bar{e}_i}\right) + \frac{2}{N}\mu_{\rm H} = 0; \tag{19.81}$$

(4) the quark and lepton Yukawa couplings give relations

$$\mu_{q_i} - \mu_\phi - \mu_{d_j} = 0, \quad \mu_{q_i} - \mu_\phi - \mu_{u_j} = 0, \quad \mu_{\ell_i} - \mu_\phi - \mu_{e_j} = 0. \tag{19.82}$$

The number of equations here is the same as the number of unknowns. Combining these, one can solve for the chemical potentials in terms of the lepton chemical potential, and finally in terms of the initial $B - L$. With N generations,

$$B = \frac{8N + 4}{22N + 13}(B - L). \tag{19.83}$$

Reasonable values of the neutrino parameters give asymmetries of the order we seek to explain. Note sources of small numbers.

(1) The phases in the couplings.
(2) The loop factor.
(3) The small density of the N particles when they drop out of equilibrium. Parametrically, one has, e.g., for production,

$$\Gamma \sim e^{(-M/T)} g^2 T \tag{19.84}$$

which is much less than $H \sim T^2/M_{\rm p}$ once the density is suppressed by $T/M_{\rm p}$, i.e. of order 10^{-6} for a 10^{13} GeV particle.

It is interesting to ask: assuming that these processes are the source of the observed asymmetry, how many parameters which enter into the computation can be measured? It is likely that, over time, many of the parameters of the light neutrino mass matrices, including possible CP-violating phases, will be measured. But while these measurements determine some of the N_i couplings and masses, they are not, in general, enough. In order to give a precise calculation, analogous to the calculations of nucleosynthesis, of the baryon number density, one needs additional information about the masses of the fields N_i. One either requires some other (currently unforeseen) experimental access to this higher-scale physics, or a compelling theory of neutrino mass in which symmetries, perhaps, reduce the number of parameters.

19.5.4 Baryogenesis through coherent scalar fields

In supersymmetric theories, the ordinary quarks and leptons are accompanied by scalar fields. These scalar fields carry baryon and lepton number. A coherent field, i.e. a large classical value of such a field, can in principle carry a large amount of baryon number. As we will see, it is quite plausible that such fields were excited in the early universe, and this can lead to a baryon asymmetry.

To understand the basics of the mechanism, consider first a model with a single complex scalar field. Take the Lagrangian to be

$$\mathcal{L} = |\partial_\mu \phi|^2 - m^2 |\phi|^2. \tag{19.85}$$

This Lagrangian has a symmetry, $\phi \to e^{i\alpha}\phi$, and a corresponding conserved current, which we will refer to as baryon number:

$$j^\mu_{\rm B} = i(\phi^* \partial^\mu \phi - \phi \partial^\mu \phi^*). \tag{19.86}$$

It also possesses a "CP" symmetry:

$$\phi \leftrightarrow \phi^*. \tag{19.87}$$

With supersymmetry in mind, we will think of m as of order M_W.

If we focus on the behavior of spatially constant fields, $\phi(\vec{x}, t) = \phi(t)$, this system is equivalent to an isotropic harmonic oscillator in two dimensions. In field theory, however, we expect that higher-dimension terms will break the symmetry. In the isotropic oscillator analogy, this corresponds to anharmonic terms which break the rotational invariance. With a general initial condition, the system will develop some non-zero angular momentum. If the motion is damped, so that the amplitude of the oscillations decreases, these rotationally non-invariant terms will become less important with time.

In the supersymmetric field theories of interest, supersymmetry will be broken by small quartic and higher-order couplings, as well as by soft masses for the scalars. So as a simple model, take:

$$\mathcal{L}_I = \lambda|\phi|^4 + \epsilon\, \phi^3\phi^* + \sigma\, \phi^4 + \text{c.c.} \tag{19.88}$$

These interactions clearly violate "B." For general complex ϵ and σ, they also violate CP. As we will shortly see, once supersymmetry is broken, quartic and higher-order couplings can be generated, but these couplings $\lambda, \epsilon, \sigma \ldots$ will be extremely small, $\mathcal{O}(m_{3/2}^2/M_{\rm p}^2)$ or $\mathcal{O}(m_{3/2}^2/M_{\rm GUT}^2)$.

In order that these tiny couplings lead to an appreciable baryon number, it is necessary that the fields, at some stage, were very large. To see how the cosmic evolution of this system can lead to a non-zero baryon number, first note that at very early times, when the Hubble constant, $H \gg m$, the mass of the field is irrelevant. It is thus reasonable to suppose that at this early time $\phi = \phi_0 \gg 0$; later we will describe some specific suggestions as to how this might come about. This system then evolves like the axion or moduli. In the the radiation and matter dominated eras, respectively, one has that

$$\phi = \frac{\phi_0}{(mt)^{3/2}} \sin(mt) \ \text{(radiation)} \tag{19.89}$$

$$= \frac{\phi_0}{(mt)} \sin(mt) \ \text{(matter)}. \tag{19.90}$$

In either case, the energy behaves, in terms of the scale factor, $R(t)$, as

$$E \approx m^2\phi_0^2 \left(\frac{R_0}{R}\right)^3, \tag{19.91}$$

i.e. it decreases like R^3, as would the energy of pressureless dust. One can think of this oscillating field as a coherent state of ϕ particles with $\vec{p} = 0$.

Now let's consider the effects of the quartic couplings. Since the field amplitude damps with time, their significance will decrease with time. Suppose, initially, that $\phi = \phi_0$ is real. Then the imaginary part of ϕ satisfies, in the approximation that ϵ and δ are small,

$$\ddot{\phi}_i + 3H\dot{\phi}_i + m^2\phi_i \approx \text{Im}(\epsilon + \delta)\phi_{\rm r}^3. \tag{19.92}$$

For large times, the right-hand side falls as $t^{-9/2}$, whereas the left-hand side falls off only as $t^{-3/2}$. As a result, just as in our mechanical analogy, baryon number (angular momentum) violation becomes negligible. The equation goes over to the free equation, with a solution of the form

$$\phi_i = a_{\rm r}\frac{{\rm Im}(\epsilon+\delta)\phi_0^3}{m^2(mt)^{3/4}}\sin(mt+\delta_{\rm r}) \ \text{(radiation)}, \tag{19.93}$$

$$\phi_i = a_{\rm m}\frac{{\rm Im}(\epsilon+\delta)\phi_0^3}{m^3 t}\sin(mt+\delta_{\rm m}) \ \text{(matter)}, \tag{19.94}$$

in the radiation and matter dominated cases, respectively. The constants $\delta_{\rm m}$, $\delta_{\rm r}$, $a_{\rm m}$ and $a_{\rm r}$ can easily be obtained numerically, and are of order unity:

$$a_{\rm r} = 0.85 \quad a_{\rm m} = 0.85 \quad \delta_{\rm r} = -0.91 \quad \delta_{\rm m} = 1.54. \tag{19.95}$$

But now we have a non-zero baryon number; substituting in the expression for the current,

$$n_{\rm B} = 2a_{\rm r}{\rm Im}(\epsilon+\delta)\frac{\phi_0^2}{m(mt)^2}\sin(\delta_{\rm r}+\pi/8) \ \text{(radiation)} \tag{19.96}$$

$$n_{\rm B} = 2a_{\rm m}{\rm Im}(\epsilon+\delta)\frac{\phi_0^2}{m(mt)^2}\sin(\delta_{\rm m}) \ \text{(matter)}. \tag{19.97}$$

Note that CP violation here can be provided by phases in the couplings and/or the initial fields. Note also as expected, $n_{\rm B}$ is conserved at late times, in the sense that the baryon number per comoving volume is constant.

This mechanism for generating baryon number could be considered without supersymmetry. In that case, it begs several questions.

- What are the scalar fields carrying baryon number?
- Why are the ϕ^4 terms so small?
- How are the scalars in the condensate converted to more familiar particles?

In the context of supersymmetry, there is a natural answer to each of these questions. First, as we have stressed, there are scalar fields carrying baryon and lepton number. As we will see, in the limit that supersymmetry is unbroken, there are typically directions in the field space in which the quartic terms in the potential vanish. Finally, the scalar quarks and leptons will be able to decay (in a baryon- and lepton-number conserving fashion) to ordinary quarks.

19.6 Flat directions and baryogenesis

To discuss the problem of baryon number generation, we first want to examine the theory in a limit in which we ignore the soft SUSY-breaking terms. After all,

at very early times, $H \gg M_W$, and these terms are irrelevant. We are now quite familiar with the fact that supersymmetric theories often exhibit flat directions. At the renormalizable level, the MSSM possesses many flat directions. A simple example is

$$H_u = \begin{pmatrix} 0 \\ v \end{pmatrix} \quad L_1 = \begin{pmatrix} v \\ 0 \end{pmatrix}. \tag{19.98}$$

This direction is characterized by a modulus which carries lepton number. Written in a gauge-invariant fashion, $\Phi = H_u L$. As we have seen, producing a lepton number is for all intents and purposes like producing a baryon number. Nonrenormalizable, higher-dimension terms, with more fields, can lift the flat direction. For example, the quartic term in the superpotential:

$$\mathcal{L}_4 = \frac{1}{M}(H_u L)^2 \tag{19.99}$$

respects all of the gauge symmetries and is invariant under R-parity. It gives rise to a potential

$$V_{\text{lift}} = \frac{|v|^6}{M^2}. \tag{19.100}$$

There are many more flat directions, and many of these do carry baryon or lepton number. A flat direction with both baryon and lepton number excited is the following:

$$\text{First generation: } Q_1^1 = b \quad \bar{u}_2 = a \quad L_2 = b \tag{19.101}$$

$$\text{Second: } \bar{d}_1 = \sqrt{|b|^2 + |a|^2} \tag{19.102}$$

$$\text{Third: } \bar{d}_3 = a. \tag{19.103}$$

(On Q, the upper index is a color index, the lower index an $SU(2)$ index, and we have suppressed the generation indices.)

Higher-dimension operators again can lift this flat direction. In this case the leading term is:

$$\mathcal{L}_7 = \frac{1}{M^3}[Q^1 \bar{d}^2 L^1][\bar{u}^1 \bar{d}^2 \bar{d}^3]. \tag{19.104}$$

Here the superscripts denote flavor. We have suppressed color and $SU(2)$ indices, but the braces indicate sets of fields which are contracted in $SU(3)$ and $SU(2)$ invariant ways. In addition to being completely gauge-invariant, this operator is invariant under ordinary R-parity. (There are lower-dimension operators, including operators of dimension 4, which violate R-parity). It gives rise to a term in the

potential:

$$V_{\text{lift}} = \frac{\Phi^{10}}{M^6}. \tag{19.105}$$

Here Φ refers in a generic way to the fields whose vevs parameterize the flat directions (a,b).

19.7 Supersymmetry breaking in the early universe

We have indicated that higher-dimension, supersymmetric operators give rise to potentials in the flat directions. To fully understand the behavior of the fields in the early universe, we need to consider supersymmetry breaking, which gives rise to additional potential terms.

In the early universe, we expect supersymmetry is much more badly broken than it is in the present era. For example, during inflation, the non-zero energy density (cosmological constant) breaks supersymmetry. Suppose that I is the inflaton field, and that the inflaton potential arises because of a non-zero value of the auxiliary field for I, $F_I = \partial W/\partial I$. So, during inflation, supersymmetry is broken by a large amount. Not surprisingly, as a result, there can be an appreciable supersymmetry-breaking potential for Φ. These contributions to the potential have the form:

$$V_{\text{H}} = H^2\Phi^2 f\left(\Phi^2/M_{\text{p}}^2\right). \tag{19.106}$$

It is perfectly possible for the second derivative of the potential near the origin to be negative. In this case, writing our higher-dimension term as:

$$W_n = \frac{1}{M^n}\Phi^{n+3}. \tag{19.107}$$

the potential takes the form

$$V = -H^2|\Phi|^2 + \frac{1}{M^{2n}}|\Phi|^{2n+4}. \tag{19.108}$$

The minimum of the potential then lies at:

$$\Phi_0 \approx M\left(\frac{H}{M}\right)^{\frac{1}{n+1}}. \tag{19.109}$$

More generally, one can see that the higher the dimension of the operator that raises the flat direction, the larger the starting value of the field – and the larger the ultimate value of the baryon number. Typically, there is plenty of time for the field to find its minimum during inflation. After inflation, H decreases, and the field Φ evolves adiabatically, oscillating slowly about the local minimum for some time.

Our examples illustrate that in models with R-parity, the value of n, and hence the size of the initial field, can vary appreciably. Which flat direction is most important depends on the form of the mass matrix (i.e. on which directions the curvature of the potential is negative). With further symmetries, it is possible that n is larger, and even that all operators which might lift the flat direction are forbidden. For the rest of this section we will continue to assume that the flat directions are lifted by terms in the superpotential. If they are not, the required analysis is different, since the lifting of the flat direction is entirely associated with supersymmetry breaking.

19.7.1 Appearance of the baryon number

The term in the potential, $|\partial W/\partial\Phi|^2$, does not break either baryon number or CP. In most models, it turns out that the leading sources of B and CP violation come from supersymmetry-breaking terms associated with F_I. These have the form

$$am_{3/2}W + bHW. \tag{19.110}$$

Here a and b are *complex*, dimensionless constants. The relative phase in these two terms, δ, violates CP. This is crucial; if the two terms carry the same phase, then the phase can be eliminated by a field redefinition, and we have to look elsewhere for possible CP-violating effects. Examining Eqs. (19.99) and (19.104), one sees that the term proportional to W violates B and/or L. In following the evolution of the field Φ, the important era occurs when $H \sim m_{3/2}$. At this point, the phase misalignment of the two terms, along with the B-violating coupling, leads to the appearance of a baryon number. From the equations of motion, the equation for the time rate of change of the baryon number is

$$\frac{dn_{\rm B}}{dt} = \frac{\sin(\delta)m_{3/2}}{M^n}\phi^{n+3}. \tag{19.111}$$

Assuming that the relevant time is H^{-1}, one is led to the estimate (supported by numerical studies)

$$n_{\rm B} = \frac{1}{M^n}\sin(\delta)\Phi_0^{n+3}. \tag{19.112}$$

Here, Φ_0 is determined by $H \approx m_{3/2}$, i.e. $\Phi_0^{2n+2} = m_{3/2}^2 M^{2n}$.

19.8 The fate of the condensate

Of course, we don't live in a universe dominated by a coherent scalar field. In this section, we consider the fate of a homogeneous condensate, ignoring possible

inhomogeneities. The following section will deal with inhomogeneities, and the interesting array of phenomena to which they might give rise.

We will adopt the following model for inflation. The features of this picture are true of many models of inflation, but by no means all. We will suppose that the energy scale of inflation is $E \sim 10^{15}$ GeV. We assume that inflation is due to a field, the inflaton I. The amplitude of the inflaton, just after inflation, we will take to be of order $M \approx 10^{18}$ GeV (the usual reduced Planck mass). Correspondingly, we will take the mass of the inflaton to be $m_I = 10^{12}$ GeV (so that $m_I^2 M_{\rm p}^2 \approx E^4$). Correspondingly, the Hubble constant during inflation is of order $H_I \approx E^2/M_{\rm p} \approx 10^{12}$ GeV.

After inflation ends, the inflaton oscillates about the minimum of its potential, much like the field Φ, until it decays. We will suppose that the inflaton couples to ordinary particles with a rate suppressed by a single power of the Planck mass. Dimensional analysis then gives for the rough value of the inflaton lifetime:

$$\Gamma_I = \frac{m_I^3}{M^2} \sim 1 \text{ GeV}. \tag{19.113}$$

The reheating temperature can then be obtained by equating the energy density at time $H \approx \Gamma(\rho = 3H^2M^2)$ to the energy density of the final plasma:

$$T_{\rm R} = T\left(t = \Gamma_I^{-1}\right) \sim (\Gamma_I M_{\rm p})^{1/2} \sim 10^9 \,\text{GeV}. \tag{19.114}$$

The decay of the inflaton is actually not sudden, but leads to a gradual reheating of the universe, as described, for example, in the book by Kolb and Turner (1990). As a function of time (H):

$$T \approx \left(T_{\rm R}^2 H(t) M_{\rm p}\right)^{1/4}. \tag{19.115}$$

As for the field Φ, our basic assumption is that during inflation, it obtains a large value, in accord with Eq. (19.109). When inflation ends, the inflaton, by assumption, still dominates the energy density for a time, oscillating about its minimum; the universe is matter dominated during this period. The field Φ now oscillates about a time-dependent minimum, given by Eq. (19.109). The minimum decreases in value with time, dropping to zero when $H \sim m_{3/2}$. During this evolution, a baryon number develops classically. This number is frozen once $H \sim m_{3/2}$.

Eventually the condensate will decay, through a variety of processes. As we have stressed, the condensate can be thought of as a coherent state of Φ particles. These particles – linear combinations of the squark and slepton fields – are unstable and will decay. However, for $H \leq m_{3/2}$, the lifetimes of these particles are much longer than in the absence of the condensate. The reason is that the fields to which Φ couples have mass of order Φ, and Φ is large. In most cases, the most important

process which destroys the condensate is what we might call evaporation: particles in the ambient thermal bath can scatter off of the particles in the condensate, leaving final states with only ordinary particles.

We can make a crude estimate for the reaction rate as follows. Because the particles which couple directly to Φ are heavy, interactions of Φ with light particles must involve loops. So we include a loop factor in the amplitude, of order α_2^2, the weak coupling squared. Because of the large masses, the amplitude is suppressed by Φ. Squaring, and multiplying by the thermal density of scattered particles, gives:

$$\Gamma_{\rm p} \sim \alpha_2^2 \pi \frac{1}{\Phi^2} \left(T_{\rm R}^2 H M\right)^{3/4}. \tag{19.116}$$

The condensate will evaporate when this quantity is of order H. Since we know the time dependence of Φ, this allows us to solve for this time. One finds that equality occurs, in the case $n = 1$, for $H_I \sim 10^2$–10^3 GeV. For $n > 1$, it occurs significantly later (for $n < 4$, it occurs before the decay of the inflaton; for $n \geq 4$, a slightly different analysis is required than that which follows). In other words, for the case $n = 1$, the condensate evaporates shortly after the baryon number is created, but for larger n, it evaporates significantly later.

The expansion of the universe is unaffected by the condensate as long as the energy density in the condensate, $\rho_\Phi \sim m_\Phi^2 \Phi^2$, is much smaller than that of the inflaton, $\rho_I \sim H^2 M^2$. Assuming that $m_\Phi \sim m_{3/2} \sim 0.1$–$1$ TeV, a typical supersymmetry breaking scale, one can estimate the ratio of the two densities at the time when $H \sim m_{3/2}$ as

$$\frac{\rho_\Phi}{\rho_I} \sim \left(\frac{m_{3/2}}{M_{\rm p}}\right)^{2/(n+1)}. \tag{19.117}$$

We are now in a position to calculate the baryon to photon ratio in this model. Given our estimate of the inflaton lifetime, the coherent motion of the inflaton still dominates the energy density when the condensate evaporates. The baryon number is just the Φ density just before evaporation divided by the Φ mass (assumed of order $m_{3/2}$), while the inflaton number is ρ_I / M_I. So the baryon to inflaton ratio follows from Eq. (19.117). With the assumption that the inflaton energy density is converted to radiation at the reheating temperature, $T_{\rm R}$, we obtain:

$$\frac{n_{\rm B}}{n_\gamma} \sim \frac{n_{\rm B}}{(\rho_I / T_{\rm R})} \sim \frac{n_{\rm B}}{n_\Phi} \frac{T_{\rm R}}{m_\Phi} \frac{\rho_\Phi}{\rho_I} \sim 10^{-10} \left(\frac{T_{\rm R}}{10^9\ {\rm GeV}}\right) \left(\frac{M_{\rm p}}{m_{3/2}}\right)^{\frac{(n-1)}{(n+1)}}. \tag{19.118}$$

Clearly the precise result depends on factors beyond those indicated here explicitly, such as the precise mass of the Φ particle(s). But as a rough estimate, it is rather robust. For $n = 1$, it is in *precisely the right range* to explain the observed baryon

asymmetry. For larger n, it can be significantly larger. In light of our discussion of late decays of moduli this is potentially quite interesting. These decays produce a huge amount of entropy, typically increasing the entropy by a factor of 10^7 or so. The baryon density is diluted by a corresponding factor. But we see that coherent production can readily yield baryon to photon densities, prior to moduli decay, of the needed size.

There are many other issues which can be studied, both in leptogenesis and in Affleck–Dine baryogenesis, but it appears that both types of process might well account for the observed baryon asymmetry. The discovery (or not) of low-energy supersymmetry, and further studies of neutrino masses, might make one or the other picture more persuasive. Both pose challenges, as they involve couplings which we are not likely to measure directly.

19.9 Dark energy

It has long been recognized that any cosmological constant in nature is far smaller than scales of particle physics. Before the discovery of the dark energy, many physicists conjectured that for some reason of principle, this energy is zero. However, it is now clear that the energy is non-zero, and in fact that this dark energy is the largest component of the energy density of the universe. Present data is compatible with the idea that this energy density represents a cosmological constant ($w = -1$), but other suggestions, typically involving time-varying scalar fields, have been offered.

Apart from its smallness, another puzzle surrounding the cosmological constant is simply one of coincidence: why is the dark energy density today comparable to the dark matter density? Weinberg has argued that it couldn't be much different than this in a universe containing stars and galaxies, provided all of the other laws of nature are as we observe. The basic point is that if the dark energy were, say, 10^3 times more dense than we observe, it would have come to dominate the energy density when the universe was much younger than it is today – prior to the formation of galaxies and stars. The rapid acceleration after that time would have prevented the formation of structure. More refined versions of the argument give estimates for the dark energy within a factor of ten of the measured value.

Weinberg speculated that perhaps the universe is much larger than we see (our current horizon). In other regions, it has different values of the cosmological constant. Only in those regions where Λ is very small would stars – and hence observers – form. Weinberg called this possible explanation (actually prediction) of Λ the *weak anthropic argument*. We will return to this question in our studies of string theory, where we will see that such a landscape of ground states may exist.

Suggested reading

Seminal papers on inflation include that of Guth (1981), which proposed a version of inflation now often referred to as "old inflation," and those of Linde (1982) and Albrecht and Steinhardt (1982), which contain the germ of the slow roll inflation idea stressed in this work. The ideas of hybrid inflation were developed by Linde (1994); those specifically discussed here were introduced by Randall *et al*. (1996) and Berkooz *et al*. (2004). There are a number of good texts on inflation and related issues, some of which we have mentioned in the previous chapter. These include those of Dodelson (2004), Kolb and Turner (1990), and Linde (1990). Dodelson provides a particularly up to date discussion of dark matter, including more detailed calculations than those presented here, and dark energy, including surveys of observational results. For a review of axions and their cosmology and astrophysics, see Turner (1990). For more recent papers which raise questions about the cosmological axion limits, see, for example, Banks *et al*. (2003). The cosmological moduli problem, and possible solutions, were first discussed by Banks *et al*. (1994) and de Carlos *et al*. (1993). A general review of electroweak baryogenesis, including detailed discussions of phenomena at the bubble walls, appears in Cohen *et al*. (1993). A discussion of electroweak baryogenesis within the MSSM appears in Carena *et al*. (2003). A detailed review of baryogenesis appears in Buchmuller *et al*. (2005), while Enqvist and Mazumdar (2003) focuses on Affleck–Dine baryogenesis. A more comprehensive review of baryogenesis mechanisms appears in Dine and Kusenko (2003). Aspects of the cosmological constant, and especially Weinberg's anthropic prediction of Λ, are explained clearly in Weinberg (1989), with more recent additions in Vilenkin (1995) and Weinberg (2000).

Exercises

(1) Verify the slow roll conditions, Eqs. (19.12) and (19.13). Determine the number of e-foldings and the size of $\delta\rho/\rho$ as a function of $\mathcal{N}$.

(2) Work through the limits on the axion in more detail. Try to analyze the behavior of the axion energy in the high-temperature regime.

(3) Construct a discrete R-symmetry which guarantees that the $H_U L$ flat direction is exactly flat. Assuming that the universe reheats to 100 MeV when a modulus decays, estimate the final baryon number of the universe in this case.

(4) Suppose that the characteristic scale of supersymmetry breaking is much higher than 1 TeV, say 10^9 GeV. Discuss baryogenesis by coherent scalar fields in such a situation.

Part 3

String theory

20

Introduction

String theory was stumbled on, more or less, by accident. In the late 1960s, string theories were first proposed as theories of the strong interactions. But, it was quickly realized that, while hadronic physics has a number of string-like features, string theories were not suitable for a detailed description. In their simplest form, string theories had massless spin-two particles and more than four dimensions of space-time, hardly features of the strong interactions. But a small group of theorists appreciated that the presence of a spin-two particle implied that these theories were generally covariant and explored them through the 1970s and and early 1980s as possible theories of quantum gravity. Like field theories, the number of possible string theories seemed to be infinite, while, unlike field theories, there was reason to believe that these theories did not suffer from ultraviolet divergences. In the 1980s, however, studies of anomalies in higher dimensions suggested that all string theories with chiral fermions and gauge interactions suffered from quantum anomalies. But in 1984, it was shown that anomalies cancel for two choices of gauge group. It was quickly recognized that the non-anomalous string theories do come close to unifying gravity and the Standard Model of particle physics. Many questions remained. Beginning in 1995, great progress was made in understanding the deeper structure of these theories. All of the known string theories were understood to be different limits of some larger structure. As string theories still provide the only framework in which one can do systematic computations of quantum gravity effects, many workers use the term "string theory" to refer to some underlying structure which unifies quantum mechanics, gravity and gauge interactions.

String theory has provided us with many insights into what a fundamental theory of gravity and gauge interactions might look like, but there is still much we do not understand. We can't begin a course by enunciating some great principle and seeing what follows. We might, for example, have imagined that the underlying theory would be a string field theory, whose basic objects would be objects which would create and annihilate strings. Some set of organizing principles would

determine the action for this system, and the rest would be a problem of working out the consequences. But there are good reasons to believe that string theory is not like this. Instead, we can at best provide a collection of facts, organized according to the teacher/author/professor's view of the subject at any given moment. As a result, it is perhaps useful first to give at least some historical perspective as to how these facts were accumulated, if only to show that there are, as of yet, no canonical texts or sacred principles in the subject. In the next section, we review a bit of the remarkable history of string theory. In the following section, we will attempt to survey what is known as of this writing: the various string theories, with their spectra and interactions, and the connections between them.

20.1 The peculiar history of string theory

For electrodynamics, the passage from classical to quantum mechanics is reasonably straightforward. But general relativity and quantum mechanics seem fundamentally incompatible. Viewed as a quantum field theory, Einstein's theory of general relativity is a non-renormalizable theory; its coupling constant (in four dimensions) has dimensions of inverse mass-squared. As a result, quantum corrections are very divergent. From the point of view developed in Part 1, these divergences should be thought of as cut off at some scale associated with new physics; general relativity is an incomplete theory. Hawking has discussed another sense in which gravity and quantum mechanics seem to clash. Hawking's paradox appears to be associated with phenomena at arbitrarily large distances – in particular, with the event horizons of large black holes. Because black holes emit a thermal spectrum of radiation, it seems possible for a pure state – a large black hole – to evolve into a mixed state. These puzzles suggest that reconciling quantum mechanics and gravity will require a radical rethinking of our understanding of very short-distance physics.

Apart from its potential to reconcile quantum mechanics and general relativity, there is another reason that string theory has attracted so much attention: it is finite, free of the ultraviolet divergences that plague ordinary quantum field theories. In the previous chapters of this book, we have adopted the point of view that our theories of nature should be viewed as effective theories. It is not clear that they can be complete in any sense. One might wonder whether some sort of structure exists where the process stops; where some finite, fundamental theory accounts for the features of our present, more tentative constructions. Many physicists have speculated through the years that these two questions are related; that an understanding of quantum general relativity would provide a fundamental length scale. The finiteness of string theory suggests it might play this role.

String theory was discovered, by accident, in the 1960s, as physicists tried to understand certain regularities of the hadronic S-matrix. In particular, hadronic scattering amplitudes exhibited a feature then referred to as "duality." Scattering amplitudes with two incoming and two outgoing particles (so-called $2 \to 2$ processes) could be described equally well by an exchange of mesons in the s channel or in the t channel (but not both simultaneously). This is not a property, at least perturbatively, of conventional quantum field theories. Veneziano succeeded in writing an expression for an S-matrix with just the required properties. Veneziano's result was extended in a variety of ways and it was soon recognized (by Nambu, Susskind and others) that this model was equivalent to a theory of strings.

One could well imagine coming to string theory by a different route. Quantum field theory describes point particles. Apart from properties like mass and charge, no additional features (size, shape) are assigned to the basic entities. One could well imagine that this is naive, but in describing nature, quantum field theory is extraordinarily successful. In fact, there is no evidence for any size of the electron or the quarks down to distances of order 10^{-17} cm (energy scales of order several TeV). Still, it is natural to try to go beyond the idea of particles as points. The simplest possibility is to consider entities with a one-dimensional extent, strings. In the next chapters, we will discuss the features of theories of string. Here we just note that a straightforward analysis yields some remarkable results. A relativistic quantum string theory is necessarily:

(1) a theory of general relativity;
(2) a theory with gauge interactions;
(3) finite: string world sheets are smooth. Strings do not interact at space-time points. As a result, in perturbation theory, one does not have the usual ultraviolet divergences of quantum theories of relativistic particles.

These features are not postulated; they are inevitable. Other, seemingly less desirable, features also emerge: the space-time dimension has to be 26 or 10. Many string theories also contain tachyons in their spectrum, whose interpretation is not immediately clear.

As theories of hadronic physics, string theories had only limited success. Their spectra and S-matrices did share some features in common with those of the real strong interactions. But as a result of the features described above – massless particles and unphysical space-time dimensions as well as the presence of tachyons in many cases – strings were quickly eclipsed by QCD as a theory of the strong interactions.

Despite these setbacks, string theory remained an intriguing topic. String theories were recognized to have short distance behavior much different – and better – than that of quantum field theories. There was reason to think that such theories were free

of ultraviolet divergences altogether. Scherk and Schwarz, and also Yoneya, made the bold proposal that string theories might well be sensible theories of quantum gravity. At the time, any concrete realization of this suggestion seemed to face enormous hurdles. The first string theories contained bosons only. But string theories with fermions were soon studied, and were discovered to have another remarkable, and until then totally unfamiliar, property: supersymmetry. We have already learned a great deal about supersymmetry, but at this early stage, its possible role in nature was completely unclear. In their early formulations, string theories only made sense in special – and at first sight uninteresting – space-time dimensions. But it had been conjectured since the work of Kaluza and Klein that higher-dimensional space-times might be "compactified," leaving theories which appear four-dimensional, and Scherk and Schwarz hypothesized that this might be the case for string theories. Over a decade, Green and Schwarz studied the supersymmetric string theories further, developing a set of calculational tools in which supersymmetry was manifest, and which were suitable for tree-level and one-loop computations.Witten and Alvarez-Gaume, however, pointed out that higher-dimensional theories in general suffer from anomalies, which render them inconsistent. They argued that almost all of the then-known chiral string theories suffered from just such anomalies. It appeared that the string program was doomed; only two known string theories, theories without gauge interactions, seemed to make sense. Green and Schwarz, however, persisted. By a direct string computation, they discovered that, while it was true that almost all would-be string theories with gauge symmetries are inconsistent, there was one exception among the then-known theories, with a gauge group $O(32)$. They reviewed the standard anomaly analysis and realized why $O(32)$ is special; this work raised the possibility that there might be one more consistent string theory, based on the gauge group $E_8 \times E_8$. The corresponding string theory, as well as another with gauge group $O(32)$ (known as the heterotic string theories), was promptly constructed.

This work stimulated widespread interest in string theory as a unified theory of all interactions, for now these theories appeared to be not only finite theories of gravity, but also nearly unique. Compactification of the heterotic string on six-dimensional manifolds known as Calabi–Yau spaces were quickly shown to lead to theories which at low energies closely resemble the Standard Model, with similar gauge groups, particle content, and other features such as repetitive generations, low-energy supersymmetry and dynamical supersymmetry breaking. The various string theories have since been shown to be part of a larger theory, suggesting that one is studying some unique structure which describes quantum gravity. Some of the basic questions about quantum gravity theories, such as Hawking's puzzle, have been, at least partially, resolved.

Many questions remain, however. There is still no detailed understanding of how string theory can make contact with experiment. There are a number of reasons for this. String theory, as we will see, is a theory with no dimensionless parameters. This is a promising beginning for a possible unified theory. But it is not clear how a small expansion parameter can actually emerge, allowing systematic computation. String theory provides no simple resolution of the cosmological constant puzzle. Finally, while there are solutions which resemble nature, there are vastly more which don't. A principle, or dynamics, which might explain the selection of one vacuum or another, has not emerged.

Yet string theory is the only model we have for a quantum theory of gravity. More than that, it is the only model we have for a finite theory, which could be viewed as some sort of ultimate theory. At the same time, string theory addresses almost all of the deficiencies we have seen in the Standard Model, and has the potential to encompass all of the solutions we have proposed. The following are some examples.

(1) The theory unifies gravity and gauge interactions in a consistent, quantum mechanical framework.
(2) The theory is completely finite. It has no free parameters. The constants of nature must be determined by the dynamics, or other features internal to the theory.
(3) The theory possesses solutions in which space-time is four-dimensional, with gauge groups close to the Standard Model and repetitive generations. It is in principle possible to compute the parameters of the Standard Model.
(4) Many of the solutions exhibit low-energy supersymmetry, of the sort we have considered in the first part of this book.
(5) Other solutions exhibit large dimensions, technicolor-like structures, and the like.
(6) The theory does not have continuous global symmetries, but often possesses discrete symmetries, of the sorts we have considered.

While these are certainly encouraging signs, as we will learn in the third part of this book, we are a long way from a detailed understanding of how string theory might describe nature. We will see that there are fundamental obstacles to such an understanding. At the same time, we will see that string theory provides a useful framework in which to assess proposals for beyond the Standard Model physics.

The third part of this book is intended to provide the reader with an overview of superstring theory, with a view to connecting string theory with nature. In the next chapter, we will study the bosonic string. We will understand how to find the spectra of string theories. We will also understand string interactions. The reason that string theories are so constrained is that strings can only interact in a limited set of ways, essentially by splitting and joining. We will explain how to translate this into concrete computations of scattering amplitudes.

In subsequent chapters, we will turn to the superstring theories, obtaining their spectra and understanding their interactions. We will then turn to compactification of string theories, focussing mainly on compactifications to four dimensions. We first consider toroidal compactifications of strings, whose features can be worked out quite explicitly. We also discuss orbifolds, simple string models which can exhibit varying amounts of supersymmetry. Then we devote a great deal of attention to compactifications on Calabi–Yau spaces. These are smooth spaces; superstring theories compactified on these spaces exhibit varying amounts of supersymmetry. Many look quite close to the real world.

Finally, we will turn to the question of developing a realistic string phenomenology. Having seen the many intriguing features of string models, we will point out some of the challenges. These are as follows.

(1) The proliferation of classes of string vacua.
(2) Within different classes, the existence of moduli.
(3) Mechanisms which generate potentials for moduli are known, but in regimes where calculations can be performed systematically, tend not to produce stable minima. The question of supersymmetry breaking is closely related to the question of stabilizing moduli.
(4) There are detailed issues, such as proton decay, features of quark and lepton masses, and many others.

We will touch on some proposed solutions to these puzzles. Much string model building simply posits that moduli have been fixed in some way, and a vacuum with desirable properties is somehow selected by some (unknown) overarching principle. This is often backed up by calculations, which, while not systematic, are at least suggestive that moduli are stabilized. An alternative viewpoint is provided by the "landscape." Here, one starts with the observation that introducing fluxes for various tensor fields can potentially stabilize moduli. The possible choices of flux vastly increase the possible array of (metastable) string ground states. If one simply accepts that there is such a landscape of states, and that the universe samples many of these states in some way, then one is led to think about distributions of parameters of low-energy physics, not merely the coupling constants, but the gauge groups, particle content, scale of supersymmetry breaking, and value of the cosmological constant. For better or worse, this is in some sense the ultimate realization of the notions of naturalness which so concerned us in Part 1. The question is why we are the likely outcome of a distribution of this sort. We will leave it for the readers – and for experiment – to sort out which, if any, of these viewpoints may be correct.

This is not a string theory textbook. The reader will not emerge from these few chapters with the level of technical proficiency in weakly coupled strings provided by Polchinski's text, or with the expertise in Calabi–Yau spaces provided by the

book of Green, Schwarz and Witten (1987). In order to quickly obtain the spectra of various string theories, the following chapters heavily emphasize light cone techniques. While some aspects of the covariant treatment are developed in order to explain the rules for computing the S-matrix, many important topics, especially the Polyakov path integral approach and BRST quantization, are given only cursory treatment. Similarly, the introduction to D-brane physics provides some basic tools, but does not touch on much of the well-developed machinery of the subject. The reader who wishes a more thorough grounding in the physics of D-branes will want to consult the texts of Polchinski (1998) and of Johnson (2003).

Suggested reading

The introduction of the book by Green *et al*. (1987) provides a particularly good overview of the history of string theory, and some of its basic structure. The introductory chapter of Polchinski's text (1998) provides a good introduction to more recent developments, and a perspective on why strings might be important in the description of nature.

21

The bosonic string

Consider a particle moving through space. As it moves, it sweeps out a path called a world line. The action of the particle is just the integral of the invariant length element along the path, up to a constant.

Suppose we want to describe the motion of a string. A string, as it moves, sweeps out a two-dimensional surface in space-time called a world sheet. We can parameterize the path in terms of two coordinates, one time-like and one space-like, denoted σ and τ, or σ_0 and σ_1. The action should not depend on the coordinates we use to parameterize the surface. Polyakov stressed that this can be achieved by using the formalism of general relativity. Introduce a two-dimensional metric, $\gamma_{\alpha\beta}$. Then an invariant action is:

$$S = \frac{T}{2} \int d^2\sigma \sqrt{-\gamma} \gamma^{\alpha\beta} \partial_\alpha X^\mu \partial_\beta X^\nu \eta_{\mu\nu}. \tag{21.1}$$

Here our conventions are such that for a flat space,

$$\gamma = \eta = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \tag{21.2}$$

(similarly, our D-dimensional space-time metric is $ds^2 = -dt^2 + d\vec{x}^2$).

This action has a large symmetry group. There are, first, general coordinate transformations of the two-dimensional surface. For a simple topology (plane or sphere), these permit us to bring the metric to the form:

$$\gamma = e^\phi \eta. \tag{21.3}$$

In this gauge (the conformal gauge) the action is independent of ϕ:

$$S = \frac{-T}{2} \int d^2\sigma \eta^{\alpha\beta} \partial_\alpha X^\mu \partial_\beta X^\nu \eta_{\mu\nu}. \tag{21.4}$$

It is possible to fix this symmetry further. To motivate this gauge choice, we consider an analogous problem in field theory. In a gauge theory like QED, we can

fix a covariant gauge, $\partial \cdot A = 0$. This gauge fixing, while manifestly Lorentz invariant, is not manifestly unitary. We might try to quantize covariantly by introducing creation and annihilation operators, a^μ. These would obey:

$$[a^\mu, a^{\dagger\,\nu}] = g^{\mu\nu} \tag{21.5}$$

so that some states seem to have negative norm. If one proceeds in this way, it is necessary to prove that states with negative (or vanishing) norm can't be produced in scattering amplitudes.

One way to deal with this is to choose a non-covariant gauge. Coulomb gauge is a familiar example, but a particularly useful description of gauge theories is obtained by choosing the "light cone gauge." First, define light cone coordinates:

$$x^\pm = \frac{1}{\sqrt{2}}(x^0 \pm x^{D-1}). \tag{21.6}$$

The remaining, transverse coordinates, we will simply denote as $\vec{X}$. Correspondingly, one defines the light cone momenta:

$$p^\pm = \frac{1}{\sqrt{2}}(p^0 \pm p^{D-1}), \quad \vec{p}. \tag{21.7}$$

Note that

$$A \cdot B = -(A^+ B^- + A^- B^+) + \vec{A} \cdot \vec{B}. \tag{21.8}$$

Now we will think of x^+ as our time variable. The "Hamiltonian" generates translations in x^+; it is in fact p^-. Note that for a particle:

$$p^2 = -2p^+ p^- + \vec{p}^2. \tag{21.9}$$

The Hamiltonian is:

$$H = \frac{1}{p^+} p. \tag{21.10}$$

Having made this choice of variables, one can then make the gauge choice $A^+ = 0$. In the Lagrangian, there are no terms involving $\partial_+ A^-$, so A^- is not a dynamical field; only the $D-2$ A^is are dynamical. So we have the correct number of physical degrees of freedom. One simply solves for A^- by its equations of motion. In the early days of QCD, this description proved very useful in understanding very-high-energy scattering. In practice, similar algebraic gauges are still very useful.

Light cone coordinates, more generally, are very helpful for identifying physical degrees of freedom. Consider the problem of counting the degrees of freedom associated with some tensor field, $A^{\mu\nu\rho}$. For a massive field, one counts by going to the rest frame, and restricting the indices μ, ν, ρ to be $D-1$ dimensional. For a massless field, the relevant group is the "little group" of the Lorentz group, $SO(D-2)$. Correspondingly, one restricts the indices to be $D-2$ dimensional. So, for example,

for a massless vector, there are $D-2$ degrees of freedom. For a symmetric, traceless tensor (the graviton), there are $[(D-2)(D-1)/2]-1$. Light cone coordinates, and light cone gauge, provide an immediate realization of this counting.

For many questions in quantum field theory, covariant methods are much more powerful than the light cone. Quantum field theorists are familiar with techniques for coping with covariant gauges. These involve introduction of additional fictitious degrees of freedom (Faddeev–Popov ghosts). It is probably fair to say that most do not know much about gauges like the light cone gauge (there is almost no treatment of these topics in standard texts). But we will see in string theory that the light cone gauge is quite useful in isolating the physical degrees of freedom of strings. It lacks some of the elegance of covariant treatments, but it avoids the need to introduce an intricate ghost structure, and, as in field theory, the physical degrees of freedom are manifest. The differences between the covariant and light cone treatments, as we will see, are most dramatic when we consider supersymmetric strings. In the light cone approach of Green and Schwarz, space-time supersymmetry is manifest. In the covariant approach, it is not at all apparent. On the other hand, for the discussion of interactions, the light cone treatment tends to be rather awkward. In this chapter, we will first introduce the light cone gauge, and then go on to discuss aspects of the covariant formulation. The suggested readings will satiate the reader interested in more details of the covariant treatment.

21.1 The light cone gauge in string theory

21.1.1 Open strings

In the conformal gauge, we can use our coordinate freedom to choose $X^+ = \tau$. We also can choose the coordinates such that the momentum density, $\mathcal{P}^+$, is constant on the string. In this gauge, in D dimensions, the independent degrees of freedom of a single string are the coordinates, $X^I(\sigma, \tau)$, $I = 1, \ldots, D-2$. They are each described by the Lagrangian of a free two-dimensional field:

$$S = \frac{T}{2} \int d^2\sigma ((\partial_\tau X^I)^2 - (\partial_\sigma X^I)^2). \tag{21.11}$$

It is customary to define another quantity, α' (the "Regge slope"), with dimensions of length-squared:

$$\alpha' = \frac{1}{2\pi T}. \tag{21.12}$$

We will generally take a step further and use units with $\alpha' = 1/2$. In this case, the action is:

$$S = \frac{-1}{2\pi} \int d^2\sigma ((\partial_\tau X^I)^2 - (\partial_\sigma X^I)^2). \tag{21.13}$$

The reader should be alerted that there is another common choice of units: $\alpha' = 2$, and we will encounter this later. In this case, the action has a $1/8\pi$ out front.

In order to write equations of motion, we need to specify boundary conditions in σ. Consider, first, open strings, i.e. strings with two free ends. We want to choose boundary conditions so that when we vary the action we can ignore surface terms. There are two possible choices.

(1) Neumann boundary conditions,

$$\partial_\sigma X^I(\tau, 0) = \partial_\sigma X^I(\tau, \pi) = 0. \tag{21.14}$$

(2) Dirichlet boundary conditions:

$$X^I(\tau, 0) = X^I(\tau, \pi) = \text{ constant}. \tag{21.15}$$

It is tempting to discard the second possibility, as it appears to violate translation invariance. For now, we consider only Neumann boundary conditions, but we will return later to Dirichlet.

We want to write a Fourier expansion for the X^Is. The normalization of the coefficients is conventionally taken to be somewhat different than that of relativistic quantum field theories:

$$X^I = x^I + p^I \tau + i \sum_{n \neq 0} \frac{1}{n} \alpha_n^I e^{-in\tau} \cos(n\sigma). \tag{21.16}$$

The α_n^μs are, up to constants, ordinary creation and annihilation operators:

$$\alpha_n^I = \sqrt{n} a_n \quad \alpha_{-n}^I = \sqrt{n} a_n^\dagger. \tag{21.17}$$

Because we are working at finite volume (in the two-dimensional sense) there are normalizable zero modes, the x^Is and p^Is. They correspond to the coordinate and momentum of the center of mass of the string. From our experience in field theory, we know how to quantize this system. We impose the commutation relations:

$$[\partial_\tau X^I(\sigma, \tau), X^J(\sigma', \tau)] = \frac{-i}{\pi} \delta^{IJ}(\sigma - \sigma'). \tag{21.18}$$

This is satisfied by:

$$[x^I, p^J] = i\delta^{IJ} \quad \left[\alpha_n^I, \alpha_{n'}^J\right] = n\delta_{n+n',0}\delta^{IJ}. \tag{21.19}$$

The states of this theory can be labeled by their transverse momenta, $\vec{p}$, and by integers corresponding to the occupation numbers of the infinite set of oscillator modes. It is helpful to keep in mind that this is just the quantization of a set of free, two-dimensional fields in a finite volume.

We can write a Hamiltonian for this system. Normal ordering:

$$H = \vec{p}^2 + N + a \tag{21.20}$$

where

$$N = \sum_{n=1}^{\infty} \alpha^I_{-n} \alpha^I_n, \tag{21.21}$$

and a is a normal ordering constant. States can be labelled by the occupation numbers for each mode, N_{n_i} and the momentum, p^I,

$$|p^I, \{N_{n_i}\}\rangle \tag{21.22}$$

The light cone Hamiltonian, H, generates translations in τ. It is convenient to refine the gauge choice:

$$X^+ = p^+ \tau.$$

Since p^- is conjugate to the light cone time, x^+,

$$p^- = H/p^+, \tag{21.23}$$

or

$$p^+ p^- = \vec{p}^2 + N + a, \quad M^2 = N + a. \tag{21.24}$$

So the quantum string describes a tower of states, of arbitrarily large mass. The constant a is not arbitrary. We will see shortly that

$$a = -1. \tag{21.25}$$

This means that the lowest state is a tachyon. We can label this state simply

$$|T(\vec{p})\rangle = |\vec{p}, \{0\}\rangle \equiv |\vec{p}\rangle. \tag{21.26}$$

The state carries transverse momentum $\vec{p}$ and longitudinal momenta p^+ and p^-, and is annihilated by the infinite tower of oscillators. The significance of this instability is not immediately clear; we will close our eyes to it for now and proceed to look at other states in the spectrum. When we study the superstring, we will often find that there are no tachyons.

The first excited state is the state

$$|A^I\rangle = \alpha^I_{-1} |\vec{p}\rangle. \tag{21.27}$$

Its mass is

$$m_A^2 = 1 + a. \tag{21.28}$$

Now we can see why $a = -1$. Here, $\vec{A}$ is a vector field with $D - 2$ components. In D dimensions, a massive vector field has $D - 1$ degrees of freedom; a massless vector has $D - 2$ degrees of freedom. So $\vec{A}$ must be massless, and $a = 1$ if the theory is Lorentz invariant. Later, we will give a fancier argument for the value of a, but the content is equivalent.

At level 2, we have a number of states:

$$\alpha^I_{-2}|\vec{p}\rangle \quad \alpha^I_{-1}\alpha^J_{-1}|\vec{p}\rangle. \tag{21.29}$$

These include a vector, a scalar, and a symmetric tensor. We won't attempt here to group them into representations of the Lorentz group.

It turns out that the value of D is fixed, $D = 26$. In the light cone formulation, the issue is that the light cone theory is not manifestly Lorentz invariant. To establish that the theory is Poincaré invariant, it is necessary to construct the full set of Lorentz generators and carefully check their commutators. This analysis yields the conditions $D = 26$ and $a = -1$. We will discuss the derivation of this result a bit more later. In a manifestly covariant formulation, such as the conformal gauge, the issue is one of unitarity, as in gauge field theories. Decoupling of negative and zero norm states yields, again, the condition $D = 26$.

Turning to the gauge boson, it is natural to ask: what are the fields charged under the gauge symmetry. The answer is suggested by a picture of a meson as a quark and antiquark connected by a string. We can allow the ends of the strings to carry various types of charges. In the case of the bosonic string, these can be, for example, a fundamental and anti-fundamental of $SU(N)$. Then the string itself transforms as a tensor product of vector representations. Because the open strings include massless gauge bosons, this product must lie in the adjoint representation of the group. In the bosonic string theory, one can also have $SO(N)$ and $Sp(N)$ groups. In the case of the superstring, we will see that the group structure is highly restricted. The theory will make sense only in flat ten dimensions, and then only if the group is $O(32)$.

21.2 Closed strings

We have begun with open strings, since these are in some ways simplest, but theories of open strings by themselves are incomplete. There are always processes which will produce closed strings. For closed strings, we again have a set of fields $X^I(\sigma, \tau)$. Their action is identical to that we wrote before. But they now obey the boundary conditions:

$$X^I(\sigma + \pi, \tau) = X^I(\sigma, \tau). \tag{21.30}$$

Again, we can write a mode expansion:

$$X^I = x^I + p^I\tau + \frac{i}{2}\sum_{n\neq 0}\frac{1}{n}\left(\alpha_n^I e^{-2in(\tau-\sigma)} + \tilde{\alpha}_n^I e^{-2in(\tau+\sigma)}\right). \tag{21.31}$$

The exponential terms are the familiar solutions to the two-dimensional wave equation. One can speak of modes moving to the left ("left movers") and to the right ("right movers") on the string. Again we have commutation relations:

$$[x^I, p^J] = i\delta^{IJ} \quad \left[\alpha_n^I, \alpha_{n'}^J\right] = n\delta_{n+n'}\delta^{IJ} \quad \left[\tilde{\alpha}_n^I, \tilde{\alpha}_{n'}^J\right] = n\delta_{n+n'}\delta^{IJ}. \tag{21.32}$$

Now the Hamiltonian is:

$$H = \vec{p}^2 + N + \tilde{N} + b \tag{21.33}$$

where

$$N = \sum_{n=1}^{\infty}\alpha_{-n}^I\alpha_n^I \quad \tilde{N} = \sum_{n=1}^{\infty}\tilde{\alpha}_{-n}^I\tilde{\alpha}_n^I. \tag{21.34}$$

In working out the spectrum, there is an important constraint. There should be no special point on the string, i.e. translations in the σ direction should leave states alone. The generator of constant shifts of σ can be found by the Noether procedure:

$$\mathcal{P}_\sigma = \int d\sigma\,\partial_\tau X^I\,\partial_\sigma X^I = N - \tilde{N}. \tag{21.35}$$

So we need to impose the condition $N = \tilde{N}$ on the states.

Once more, the lowest state is a scalar,

$$|T\rangle = |\vec{p}\rangle \quad m_T^2 = b. \tag{21.36}$$

Because of the constraint, the first excited state is:

$$|\Psi_{IJ}\rangle = \tilde{\alpha}_{-1}^I\alpha_{-1}^J|\vec{p}\rangle. \tag{21.37}$$

We can immediately decompose these states into irreducible representations of the little group; there is a symmetric traceless tensor; a scalar (the trace); and an antisymmetric tensor. A symmetric, traceless tensor should have, if massive, $D^2 - D - 1$ states. Here, however, we have only $D^2 - 3D + 1$ states. This is precisely the correct number of states for a massless, spin-two particle – a graviton. The remaining states are precisely the number for a massless antisymmetric tensor field and a scalar. So we learn that $b = -2$.

This is a remarkable result. General arguments, going back to Feynman, Weinberg and others, show that a massless spin-two particle, in a relativistic theory, necessarily couples like a graviton in Einstein's theory. So string theory is a theory

of general relativity. This bosonic string is clearly unrealistic. But the presence of the graviton will be a feature of all string theories, including the more realistic ones.

21.3 String interactions

The light cone formulation is very useful for determining the spectrum of string theories, but it is somewhat more awkward for the discussion of interactions. As explained in the introduction, string interactions are determined geometrically, by the nature of the string world sheet. Actually turning drawings of world sheets into a practical computational method is surprisingly straightforward. This is most easily done using the conformal symmetry of the string theory. So we return to the conformal gauge. There are close similarities between the treatment of both open and closed strings. We will start with the treatment of closed strings, for which the Green functions are somewhat simpler. At the end of this chapter we will return to open strings.

21.3.1 String theory in conformal gauge

In conformal gauge, the action is:

$$S = \frac{1}{\pi}\int d^2\sigma((\partial_\tau X^\mu)^2 - (\partial_\sigma X^\mu)^2). \tag{21.38}$$

Introducing the two-dimensional light cone coordinates

$$\sigma_\pm = \sigma_0 \pm \sigma_1 \tag{21.39}$$

the flat world sheet metric takes the form

$$\eta_{+-} = \eta_{-+} = -\frac{1}{2} \tag{21.40}$$

and the action can be written:

$$S = \frac{1}{8\pi}\int d\sigma_+\, d\sigma_- \partial_{\sigma_+} X^\mu \partial_{\sigma_-} X^\mu. \tag{21.41}$$

At the classical level, this action is invariant under a conformal rescaling of the coordinates. If we introduce light cone coordinates on the world sheet, then the action is invariant under the transformations:

$$\sigma_\pm \to f_\pm(\sigma_\pm). \tag{21.42}$$

Later, we will Wick rotate, and work with complex coordinates; these conformal transformations will then be the conformal transformations familiar in complex variable theory. It is well known that by a conformal transformation one can map

the plane into a sphere, for example. In this case, the regions at infinity with incoming or outgoing strings are mapped to points. The creation or destruction of strings at these points is described by local operators in the two-dimensional, world sheet theory. In order to respect the conformal symmetry, these operators must, like the action, be integrals over the world sheet of local dimension-two operators. These operators are known as vertex operators, $V(\sigma, \tau)$.

In conformal gauge, the action also contains Faddeev–Popov ghost terms, associated with fixing the world sheet general coordinate invariance. We will discuss some of their features later. But let's focus on the fields X^μ first. If we simply write mode expansions for the fields (taking closed strings, for definiteness)

$$X^\mu = x^\mu + p^\mu \tau + i \sum_{n \neq 0} \frac{1}{n} \left(\alpha_n^\mu e^{-2in(\tau - \sigma)} + \tilde{\alpha}_n^\mu e^{-2in(\tau + \sigma)} \right) \tag{21.43}$$

we will encounter difficulties. The α^μs will now obey commutation relations:

$$[x^\mu, p^\nu] = i g^{\mu\nu} \quad \left[\alpha_n^\mu, \alpha_{n'}^\nu \right] = n \delta_{n+n'} g^{\mu\nu}. \tag{21.44}$$

If we proceed naively, the minus sign means from g^{00} that we will have states in the spectrum of negative or zero norm.

The appearance of negative norm states is familiar in gauge field theory. The resolution of the problem, there, is gauge invariance. One can either choose a gauge in which there are no states with negative norm, or one can work in a covariant gauge in which the negative norm states are projected out. In a modern language, this projection is implemented by the BRST procedure. But it is not hard to check that in a covariant gauge, low-order diagrams in QED, for example, give vanishing amplitudes to produce negative or zero norm states (photons with time-like or light-like polarization vectors). In gauge theories, it is precisely the gauge symmetry which accounts for this. In string theory, it is another symmetry, the residual conformal symmetry of the conformal gauge.

In the chapter on general relativity, we learned that differentiation of the matter action with respect to the metric gives the energy-momentum tensor. In Einstein's theory, differentiating the Einstein term as well gives Einstein's equations. In the string case, the world sheet metric has no dynamics (the Einstein action in two dimensions is a total derivative), and the Euler–Lagrange equation for γ yields the equation that the energy-momentum tensor vanishes. Quantum mechanically, these will be constraint equations. The components of the energy-momentum tensor are:

$$T_{10} = T_{01} = \partial_0 X \cdot \partial_1 X \quad T_{00} = T_{11} = \frac{1}{2}((\partial_0 X)^2 + (\partial_1 X)^2). \tag{21.45}$$

The energy-momentum tensor is traceless. This is a consequence of conformal invariance; you can show that the trace is the generator of conformal transformations.

In terms of the light cone coordinates, the non-vanishing components of the stress tensor are:

$$T_{++} = \partial_+ X \cdot \partial_+ X \quad T_{--} = \partial_- X \cdot \partial_- X. \tag{21.46}$$

Note that $T_{+-} = T_{-+} = 0$. Energy-momentum conservation then says:

$$\partial_- T_{++} = 0 \quad \partial_+ T_{--} = 0. \tag{21.47}$$

As a result, any quantity of the form $f(x^+)T_{++}$ or $f(x^-)T_{--}$ is also conserved. Integrating over the world sheet, this gives an infinite number of conserved charges.

We want to impose the condition of vanishing stress tensor as a condition on states. There is a problem, however, and this is one way of understanding the origin of the critical dimension, 26. The obstacle is an anomaly, similar to anomalies we encountered in the first part of this text. One can see the problem if one takes the mode expansions for the X^μs and works out the commutators for the Ts. We will show in the next section that

$$[T_{++}(\sigma), T_{++}(\sigma')] = \frac{i}{24}(26 - D)\delta'''(\sigma - \sigma') + i(T_{++}(\sigma) + T_{++}(\sigma'))\delta'(\sigma - \sigma') \tag{21.48}$$

and a similar equation for T_{--}. The first term is clearly an obstruction to imposing the constraint, unless $D = 26$. The 26 arises from the energy-momentum tensor of the Faddeev–Popov ghosts. Were it not for the ghosts, strings would *never* make sense quantum mechanically. One can calculate this commutator painstakingly by decomposing in modes. But there are simpler methods, which also provide important insights into string theory, which we develop in the next section.

21.4 Conformal invariance

The analysis of conformal invariance is enormously simplified by passing to Euclidean space. Define:

$$w = \tau + i\sigma \quad \bar{w} = \tau - i\sigma. \tag{21.49}$$

The ws describe a cylinder. Again, in this section $\alpha' = 2$. This choice will make the coordinate space Green functions for the X^μs very simple. The Euclidean action is now:

$$S = \frac{1}{8\pi} \int d^2w \; \partial_w X^\mu \partial_{\bar{w}} X^\mu. \tag{21.50}$$

In complex coordinates, the non-vanishing components of the energy-momentum tensor are:

$$T_{ww} = -\partial_w X \cdot \partial_w X \quad T_{\bar{w}\bar{w}} = -\partial_{\bar{w}} X \cdot \partial_{\bar{w}} X. \tag{21.51}$$

We saw in the previous section that the string action, in Minkowski coordinates, is invariant under the transformations:

$$\sigma^+ \to f(\sigma^+) \quad \sigma_- \to g(\sigma^-). \tag{21.52}$$

In terms of the complex coordinates, this becomes invariance under the transformations:

$$w \to f(w) \quad \bar{w} \to f^*(\bar{w}). \tag{21.53}$$

These are conformal transformations of the complex variable, and, as a result of this symmetry, we are able to bring all of the machinery of complex analysis to bear on this problem. One particularly useful conformal transformation is the mapping of the cylinder onto the complex plane:

$$z = e^w \quad \bar{z} = e^{\bar{w}}. \tag{21.54}$$

Under this mapping, surfaces of constant τ on the cylinder are mapped into circles in the complex plane; $\tau \to -\infty$ is mapped into the origin, and $\tau \to \infty$ is mapped to ∞. Surfaces of constant τ are mapped into circles.

It is convenient to write our previous expression for X^μ in terms of the variable z. First write our previous expressions:

$$\begin{aligned} X^\mu &= x^\mu + p^\mu \tau + i \sum_{n \neq 0} \frac{1}{n} \left(\alpha_n^\mu e^{-2in(\tau-\sigma)} + \tilde{\alpha}_n^\mu e^{-2in(\tau+\sigma)} \right) \\ &= X_{\mathrm{L}}^\mu + X_{\mathrm{R}}^\mu, \end{aligned} \tag{21.55}$$

where

$$X_{\mathrm{L}} = \frac{1}{2} x^\mu + \frac{1}{2} p^\mu (\tau - \sigma) + i \sum_{n \neq 0} \frac{1}{n} \alpha_n^\mu e^{-in(\tau-\sigma)} \tag{21.56}$$

$$X_{\mathrm{R}} = \frac{1}{2} x^\mu + \frac{1}{2} p^\mu (\tau + \sigma) + i \sum_{n \neq 0} \frac{1}{n} \alpha_n^\mu e^{-in(\tau+\sigma)}. \tag{21.57}$$

X_{L} is holomorphic (analytic) in z, X_{R} is antiholomorphic:

$$\partial X_{\mathrm{L}} = -i \alpha_n^\mu z^{-n-1} \quad \partial X_{\mathrm{R}} = -i \tilde{\alpha}_n^\mu \bar{z}^{-n-1}. \tag{21.58}$$

where $\alpha_0^\mu = \tilde{\alpha}_0^\mu = \frac{1}{2} p^\mu$.

Let us evaluate the propagator of the xs in coordinate space. The Xs are just two-dimensional quantum fields. Their kinetic term, however, is somewhat unconventional. Because we work with units $\alpha' = 2$, so the action has a factor of $1/8\pi$ out front. Accounting for the extra 4π, the coordinate space-propagator is (in Euclidean

space):

$$\langle X^\mu(\sigma),\, X^\nu(0)\rangle = 4\pi\,\delta^{\mu\nu}\int \frac{d^2k}{(2\pi)^2}\,\frac{e^{i\sigma\cdot k}}{k^2}. \tag{21.59}$$

This equation is logarithmically divergent in the infrared. We can use this to our advantage, cutting off the integral at scale μ and isolating the $\ln(\mu|z-z'|)$. The logarithmic dependence can be seen almost by inspection of the integral:

$$\langle X^\mu(z)X^\nu(z')\rangle = 2\delta^{\mu\nu}\ln(|z-z'|\mu) = \ln(z-z') + \ln(\bar{z}-\bar{z}') + \ln(\mu^2). \tag{21.60}$$

As we will see shortly, the infrared cutoff will drop out of physically interesting quantities, so we will suppress it in the following.

In the covariant formulation, conformal invariance is crucial to the quantum theory of strings. To understand the workings of two-dimensional conformal invariance, we can use techniques of complex variable theory, and the operator product expansion (OPE). We have discussed the operator product expansion previously, in the context of two-dimensional gauge anomalies. It is also important in QCD, in the analysis of various short-distance phenomena. The basic idea is that if one has two operators, $\mathcal{O}(z_1)$ and $\mathcal{O}(z_2)$, when $z_1 \to z_2$, we have

$$\mathcal{O}_i(z_1)\mathcal{O}_j(z_2) \underset{z_1\to z_2}{\to} \sum_k C_{ijk}(z_1-z_2)\mathcal{O}_k(z_1). \tag{21.61}$$

The coefficients C_{ijk} are, in general, singular as $z_1 \to z_2$. The power follows from dimensional analysis.

To implement this rather abstract statement, one can insert the two operators in a Green function with other operators, located at some distance from z_1. In other words, one studies:

$$\langle \mathcal{O}_i(z_1)\mathcal{O}_j(z_2)\Psi(z_3)\Psi(z_4)\ldots\rangle. \tag{21.62}$$

One can contract the operators in $\mathcal{O}(z_1)$ with those in $\mathcal{O}(z_2)$, obtaining expressions which are singular as $z_1 \to z_2$, or with the other operators, obtaining non-singular expressions. The leading term in the OPE will come from the term with the maximum number of operators at z_1 contracted with operators at z_2; less singular operators will arise when we contract fewer operators.

As an example, which will be useful in a moment, consider the product $\partial X^\mu(z)\partial X^\nu(w)$. If this appears in a Green function, the most singular term as $z \to w$ will be that where we contract $\partial X(z)$ with $\partial X(w)$. The result will be like the insertion of the unit operator at a point, times the singular function $1/(z-w)^2$, so we can write:

$$\partial X^\mu(z)\partial X^\nu(\omega) \sim \frac{g^{\mu\nu}}{(z-w)^2} + \cdots. \tag{21.63}$$

A somewhat more non-trivial, and important, set of operator product expansions are provided by the stress tensor and derivatives of X:

$$T(z)\partial X^\nu(w) = \partial X^\mu(z)\partial X^\mu(z)\partial X^\nu(w). \tag{21.64}$$

Now the most singular term arises when we contract the $\partial X(w)$ with one of the $\partial X(z)$ factors in $T(z)$. The other $\partial X(z)$ is left alone; in Green's functions, it must be contracted with other more far away operators. So we are left with:

$$T(z)\partial X(w) \approx \frac{1}{(z-w)^2}\partial X(w) + \frac{1}{z-w}\partial^2 X(w) + \cdots. \tag{21.65}$$

Another important set of operators will turn out to be exponentials of X:

$$T(z)e^{ip\cdot x} = \frac{k^2}{(z-w)^2}e^{ik\cdot x} + \cdots. \tag{21.66}$$

To get some sense of the utility of conformal invariance and OPEs, let's compute the commutators of the α^μs. Start with

$$\alpha_n^\mu = \oint \frac{dz}{2\pi} z^n \partial X^\mu, \tag{21.67}$$

where the contour is taken about the origin. Now use the fact that on the complex plane, time ordering becomes radial ordering, So, for $|z| > |w|$,

$$T\langle \partial X^\mu(z)\partial X^\nu(w)\rangle = \langle \partial X^\mu(z)\partial X^\nu(w)\rangle \tag{21.68}$$

For $|z| < |w|$,

$$T\langle \partial X^\mu(z)\partial X^\nu(w)\rangle = \langle \partial X^\nu(w)\partial X^\mu(z)\rangle. \tag{21.69}$$

Thus we have:

$$\left[\alpha_m^\mu, \alpha_n^\nu\right] = \left[\oint \frac{dz}{2\pi} z^m \oint \frac{dw}{2\pi} w^n - \oint \frac{dw}{2\pi} w^n \oint \frac{dz}{2\pi} z^m\right] \partial X^\mu(z)\partial X^\nu(w) \tag{21.70}$$

where the contour can be taken to be a circle about the origin. In the first term, we take $|z| > |w|$. In the second, $|w| > |z|$. Now to evaluate the integral, do, say, the z integral first. For fixed w, deform the z contour so that it encircles z (Fig. 21.1). Then

$$\begin{aligned}\left[\alpha_m^\mu, \alpha_n^\nu\right] &= \oint \frac{dw}{2\pi} w^n \oint \frac{dz}{2\pi} z^m \frac{1}{(z-w)^2} \\ &= m\delta_{m+n} g^{\mu\nu}.\end{aligned}$$

Let's return to the stress tensor. We expect that the stress tensor is the generator of conformal transformations, and that its commutators should contain information about the dimensions of operators. What we have just learned, by example, is that

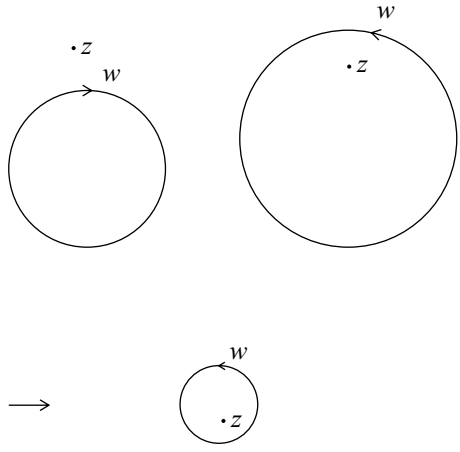

Fig. 21.1. Contour integral manipulations used to evaluate commutators in conformal field theory.

the operator products of operators encode the commutators. We could show by the Noether procedure that the stress tensor is the generator of conformal transformations. But let's simply check. Consider the transformation:

$$z \to z + \epsilon(z). \tag{21.71}$$

We expect that the generator of this transformation is:

$$\oint dz T(z)\epsilon(z). \tag{21.72}$$

We take the special case of an overall conformal rescaling:

$$\epsilon(z) = \lambda z. \tag{21.73}$$

Now suppose that we have an operator, $\mathcal{O}(w)$, and that

$$T(z)\mathcal{O}(w) = \frac{h}{(z-w)^2}\mathcal{O}(w) + \text{ less singular}. \tag{21.74}$$

Then

$$\begin{aligned}\left[\frac{1}{2\pi i}\oint T(z)\epsilon(z), \mathcal{O}(w)\right] &= \frac{1}{2\pi i}\oint dz \frac{\lambda z h \mathcal{O}(w)}{(z-w)^2} \\ &= \lambda h \mathcal{O}(w).\end{aligned} \tag{21.75}$$

This means that under the conformal rescaling, $\mathcal{O} \to h\mathcal{O}$, just as we expect for an operator of dimension h. As an example, consider $\mathcal{O} = (\partial)^n X$. This should have

dimension n, and the leading term in its OPE is just of the form of Eq. (21.74), with $h = n$.

More precisely, an operator is called a primary field of dimension d if:

$$T(z)\mathcal{O}(w) = \frac{d\mathcal{O}}{(z-w)^2} + \frac{\partial\mathcal{O}}{z-w}. \tag{21.76}$$

Note that $\partial X(z)$ is an example; $e^{ip\cdot x}$ is another. However, $(\partial)^n X$ is not, in general, as the $1/(z-w)$ term does not have quite the right form. A particularly interesting operator is the stress tensor itself. Naively, this has dimension two, but it is not a primary field. In the operator product expansion, the most singular term arises from the contraction of all of the derivative terms. This is proportional to the unit operator. The first subleading term, where one contracts just one pair of derivatives, gives a piece proportional to the stress tensor itself:

$$T(z)T(w) = \frac{D}{(z-w)^4} + \frac{1}{(z-w)^2}T(w). \tag{21.77}$$

When one includes the Faddeev–Popov ghosts, one finds that they give an additional contribution, changing D to $D-26$.

The algebra of the Fourier modes of T is known as the Virasoro algebra, and is important both in string theory, conformal field theory, and mathematics. In the string theory, it provides important constraints on states. Define the operators:

$$L_n = \frac{1}{2\pi i}\oint dz z^{n+1}T(z). \tag{21.78}$$

In terms of these,

$$T(z) = \sum_{m=-\infty}^{\infty}\frac{L_m}{z^{m+2}} \tag{21.79}$$

and similarly for $\bar{z}$. Because the stress tensor is conserved, we are free to choose any time (radius for the circle). The operator product (21.77) is equivalent to the commutation relations above. Proceeding as we did for the commutators of the αs, gives:

$$[L_n, L_m] = (m-n)L_{m+n} + \frac{D}{12}(m^3 - m)\delta_{m+n}. \tag{21.80}$$

Using expression (21.16) we can construct the L_ns:

$$L_m = \frac{1}{2}\sum : \alpha^{\mu}_{m-n}\alpha_{\mu\, n} : \quad \tilde{L}_m = \frac{1}{2}\sum : \tilde{\alpha}^{\mu}_{m-n}\tilde{\alpha}_{\mu\, n} :. \tag{21.81}$$

The colons indicates normal ordering. Only when $m = 0$ is this significant. In this case, we have to allow for the possibility of a normal ordering constant.

This constant is related to the constant we found in the Hamiltonian in light cone gauge.

$$L_0 = \sum_{n=0}^{\infty} \alpha^{\mu}_{-n}\alpha_{\mu\, n} - a \quad \tilde{L}_0 = \sum_{n=0}^{\infty} \tilde{\alpha}^{\mu}_{-n}\tilde{\alpha}_{\mu\, n} - a. \tag{21.82}$$

Now we want to consider the constraint on states corresponding to the classical vanishing of the stress tensor. Because of the commutation relations, we cannot require all of the Ls annihilate physical states. We require instead:

$$L_m|\Psi\rangle = 0 \tag{21.83}$$

for $m \geq 0$. Since $L_m^{\dagger} = L_{-m}$, this insures that

$$\langle\Psi|L_n|\Psi\rangle = 0 \; \forall \; n. \tag{21.84}$$

Because of the constraint $L_0 = \tilde{L}_0$, at the first excited level, we have the state:

$$|\epsilon\rangle = \epsilon_{\mu\nu}\alpha^{\mu}_{-1}\tilde{\alpha}^{\nu}_{-1}|p^{\mu}\rangle. \tag{21.85}$$

The L_n, for $n > 1$, trivially annihilate the state. For $n = 1$ we have:

$$L_1|\epsilon\rangle = \alpha^{\mu}_0\epsilon_{\mu\nu}|p^{\nu}\rangle. \tag{21.86}$$

Taking into account also $\tilde{L}_1$, we have the conditions:

$$p_{\mu}\epsilon^{\mu\nu} = 0 = p_{\nu}\epsilon^{\mu\nu}. \tag{21.87}$$

This is similar to the condition $k_{\mu}\epsilon^{\mu}$ familiar in covariant gauge electrodynamics and eliminates the negative norm states. Consider, now, L_0:

$$L_0|\epsilon\rangle = (p^2 - a + 1)|\epsilon\rangle. \tag{21.88}$$

So if $a = 1$, the constraint is $p^2 = 0$, as we expect from Lorentz invariance. For open strings there is an analogous construction.

21.5 Vertex operators and the S-matrix

We have argued that, when the cylinder is mapped to the plane, the creation or destruction of states is described by local operators, known as vertex operators. In this section, we discuss the properties of these operators and their construction. We explain how the space-time S-matrix is obtained from correlation functions of these operators, and compute a famous example.

21.5.1 Vertex operators

There is a close correspondence between states and operators: $z \to 0$ corresponds to $t \to -\infty$. So consider, for example,

$$\partial_z X^\mu |0\rangle, \tag{21.89}$$

as $z \to 0$. This is:

$$\partial_z X(z \to 0)|0\rangle = -i \sum_{m=-1}^{\infty} \frac{\alpha_m^\mu}{z^{m+1}} |0\rangle. \tag{21.90}$$

All but the term $m = -1$ annihilate the state to the right. Combining this with a similar left-moving operator creates a single particle state.

More generally, in conformal field theories, there is a one to one correspondence between states and operators. This is the realization of the picture discussed in the introduction. By mapping the string world sheet to the plane, the incoming/outgoing states have been mapped to points, and the production or annihilation of particles at these points is described by local operators.

The construction of the S-matrix in string theory relies on this connection of states and operators. The operators which create and annihilate states are known as vertex operators. What properties should a vertex operator possess? The production of the particle should be represented as an integral over the string world sheet (so that there is no special point along the string). The expression

$$\int d^2 z V(z, \bar{z}) \tag{21.91}$$

should be invariant under conformal transformations. This means that the operator should possess dimension two; more precisely, it should possess dimension one with respect to both the left- and the right-moving stress tensors:

$$T(z)V(w, \bar{w}) = \frac{1}{(z-w)^2} V(w, \bar{w}) + \frac{1}{z-w} \partial_w V(w, \bar{w}) + \cdots \tag{21.92}$$

and similarly for $\bar{T}$. An operator with this property is called a (1, 1) operator.

A particularly important operator in two-dimensional free field theory (i.e. the string theories we are describing up to now) is constructed from the exponential of the scalar field:

$$\mathcal{O}_p = e^{ip \cdot x}. \tag{21.93}$$

This has dimension

$$d = p^2 \tag{21.94}$$

with respect to the left-moving stress tensor, and similarly for the right-moving part.

With these ingredients, we can construct operators of dimension (1, 1). These are in one to one correspondence with the states we have found in the light cone construction.

(1) The tachyon:

$$e^{ip\cdot x} \quad p^2 = 1. \tag{21.95}$$

(2) The graviton, antisymmetric tensor, and dilaton:

$$\epsilon_{\mu\nu}\partial X^\mu \bar{\partial} X^\nu e^{ip\cdot x} \quad p^2 = 0. \tag{21.96}$$

The operator product

$$\partial X^\rho(z)\partial X_\rho(z)\epsilon_{\mu\nu}(p)\bar{\partial} X^\mu(w)\partial X^\nu(w)e^{ip\cdot x}(w) \tag{21.97}$$

contains terms which go as $1/(z-w)^3$ from contracting one of the derivatives in the stress tensor with $e^{ip\cdot x}$ and one with ∂X^μ. Examining Eq. (21.92), this leads to the requirement

$$p^\mu \epsilon_{\mu\nu}(p) = 0 \tag{21.98}$$

which we expect for massless spin-2 states. In our earlier operator discussion, this was one of the Virasoro conditions.

(3) Massive states:

$$\epsilon_{\mu_1\cdots\mu_n}(p)\partial X^{\mu_1}\partial X^{\mu_2}\cdots\partial X^{\mu_n}e^{ip\cdot x} \quad p^2 = 1 - n. \tag{21.99}$$

Obtaining the correct OPE with the stress tensor now gives a set of constraints on the polarization tensor; again these are just the Virasoro constraints. Without worrying about degeneracies, we have a formula for the masses:

$$M_n^2 = n - 1. \tag{21.100}$$

This is what we found in the light cone gauge. Traditionally, the states were organized in terms of their spins. States of a given spin all lie on straight lines, known as "Regge trajectories." These results are all in agreement with the light cone spectra we found earlier.

21.5.2 The S-matrix

Now we will make a guess as to how to construct an S-matrix. Our vertex operators, integrated over the world sheet, are invariant under reparameterizations and conformal transformation of the world sheet coordinates. We have seen that they correspond to creation and annihilation of states in the far past and far future. We

will normalize the vertex operators so that:

$$V_i(z)V_j(w) \sim \frac{1}{|z-w|^4}. \tag{21.101}$$

So we study correlation functions of the form:

$$\int d^2z_1 \dots d^2z_n \langle V_1(z_1, p_1) \dots V_n(z_n p_n)\rangle. \tag{21.102}$$

We will include a coupling constant, g, with each vertex operator.

Before evaluating this expression in special cases, let's consider the problem of evaluating correlation functions of exponentials:

$$\left\langle e^{i\sum p_i \cdot X(z_i)} \right\rangle. \tag{21.103}$$

An easy way to evaluate this expression is to work in the path integral framework. Then the exponential has the structure

$$\int d^2z J_\mu(z) X(z), \tag{21.104}$$

where

$$J_\mu(z) = \sum_i p_{i\mu} \delta^2(z - z_i). \tag{21.105}$$

But we know that the result of such a path integral is:

$$\exp\left(i \int d^2z d^2z' J_\mu(z) J^\mu(z') \Delta(z - z')\right) = \exp\left(\sum p_i \cdot p_j \ln |(z_i - z_j)|^2 \mu^2\right), \tag{21.106}$$

where we have made a point of restoring the infrared cutoff.

Let's consider the infrared cutoff first. Overall, we have a factor:

$$\mu^{(\sum p_i)^2}. \tag{21.107}$$

This vanishes as $\mu \to 0$, unless $\sum p_i = 0$, i.e. *unless momentum is conserved*. This result is related to the Mermin–Wagner–Coleman theorem that there is no spontaneous breaking of global symmetries in two dimensions. Translational invariance is a global symmetry of the two-dimensional field theory; $e^{ip\cdot x}$ transforms under the symmetry. The only non-vanishing correlation functions are translationally invariant.

This correlation function also has an ultraviolet problem, coming from the $i = j$ terms in the sum. Eliminating these corresponds to normal ordering the vertex operators, and we will do this in what follows (we can, if we like, introduce an explicit ultraviolet cutoff; this gives a factor which can be absorbed into the definition of the vertex operators).

There is one more set of divergences we need to deal with. These are associated with a part of the conformal invariance we have not yet fixed. The operators L_0, L_1 and L_{-1} form a closed algebra. On the plane, they generate overall rescalings (L_0), translations (L_1) and more general transformations (L_{-1}) which can be unified in $SL(2, C)$, the Möbius group. It acts on the world sheet coordinates as:

$$z = \frac{\alpha z' + \beta}{\gamma z' + \delta}. \tag{21.108}$$

These have the feature that they map the plane once into itself. It is necessary to fix this symmetry and divide by the volume of the corresponding gauge group. We can choose the location of three of the vertex operators, say z_1, z_2, z_3. These are conventionally taken to be 0, 1, ∞. It is necessary also to divide by the volume of this group; the corresponding factor is

$$\Omega_M = |z_1 - z_2|^2 |z_1 - z_3|^2 |z_2 - z_3|^2. \tag{21.109}$$

One can simply accept that this emerges from a Faddeev–Popov condition, or derive this following the exercises at the end of the chapter. Finally, it is necessary to divide by g_s^2. This insures that a three-particle process is proportional to g_s, a four particle amplitude to g_s^2, and so on.

Using these results we can construct particular scattering amplitudes. While physically somewhat uninteresting, the easiest to examine is simply the scattering of tachyons. Let's specialize to the case of two incoming, two outgoing particles. Putting together our results above we have (remembering that $z_3 \to \infty$):

$$\mathcal{A} = \frac{1}{\Omega_M} \int d^2 z_4 |z_1 - z_2|^2 |z_1 - z_3|^2 |z_2 - z_3|^2 |z_3|^{p_3 \cdot (p_1 + p_2 + p_3)} |z_1 - z_2|^{p_1 \cdot p_2} |z_4|^{p_4 \cdot p_1} |z_4 - 1|^{p_4 \cdot p_2}. \tag{21.110}$$

Using momentum conservation, the z_4 independent pieces cancel out in the limit, and we are left with:

$$\mathcal{A} = \int d^2 z |z|^{2 p_1 \cdot p_4} |z - 1|^{2 p_2 \cdot p_4}. \tag{21.111}$$

Now we need an integral table:

$$\begin{aligned} I &= \int |z|^{-A} |1 - z|^{-B} d^2 z \\ &= B\left(1 - \frac{A}{2}, 1 - \frac{B}{2}, \frac{A + B}{2} - 1\right). \end{aligned} \tag{21.112}$$

Here B is:

$$B = \pi \frac{\Gamma(a)\Gamma(b)\Gamma(c)}{\Gamma(a + b)\Gamma(b + c)\Gamma(c + a)}. \tag{21.113}$$

We can express this result in terms of the Mandelstam invariants for $2 \to 2$ scattering, $s = -(p_1 + p_2)^2$, $t = -(p_2 - p_3)^2$ and $u = -(p_1 - p_4)^2$. Using the mass shell conditions,

$$p_4 \cdot p_1 = \frac{1}{2}\left(u + \left(p_1^2 - p_4^2\right)\right) \quad p_4 \cdot p_2 = -(p_3 + p_2 + p_1) \cdot p_2 = \frac{1}{2}(-s - t + 2m^2), \tag{21.114}$$

gives

$$\mathcal{A} = \frac{\kappa^2}{4\pi} B(-4s + 1, -4t + 1, -4u + 1). \tag{21.115}$$

This is the Virasoro–Shapiro amplitude. There are a number of interesting features of this amplitude. It has singularities at precisely the locations of the masses of the string states. It should be noted, also, that we have obtained this result by an analytic continuation. The original integral is only convergent for a range of momenta, corresponding, essentially, to sitting "below" the threshold for the tachyon in the intermediate states.

We will not develop the machinery of open string amplitudes here, but it is similar. One again needs to compute correlation functions of vertex operators. The vertex operators are somewhat different. Also, the boundary conditions for the two-dimensional fields, and thus the Green functions, are also different. The scattering amplitude for open string tachyons is known as the Veneziano formula.

21.5.3 Factorization

The appearance of poles in the S-matrix at the masses of the string states is no accident. We can understand it in terms of our vertex operator and OPE analysis. Suppose that particles one and two, with momenta p_1 and p_2, have $s = (p_1 + p_2)^2 = -m_n^2$, the mass-squared of one of the physical states of the system. Consider the OPE of their vertex operators:

$$e^{ip_1 \cdot X(z_1)} e^{ip_2 \cdot X(z_2)} \approx e^{i(p_1 + p_2) \cdot X(z_2)} |z_1 - z_2|^{2p_1 \cdot p_2}. \tag{21.116}$$

So in the S-matrix, fixing $z_2 = 0$, $z_3 = 1$, and $z_4 = \infty$, we encounter:

$$\int d^2 z |z_1|^{2p_1 \cdot p_2} \left\langle e^{i(p_1 + p_2) \cdot X(z_2)} e^{ip_3 \cdot X(z_3)} e^{ip_4 \cdot X(z_4)} \right\rangle. \tag{21.117}$$

Using momentum conservation and the on-shell conditions for p_1 and p_2:

$$2p_2 \cdot p_1 = q^2 - 8 \tag{21.118}$$

where $q = p_1 + p_2$. So the z integral gives a pole,

$$\mathcal{A} \sim \frac{1}{4 - q^2} \tag{21.119}$$

i.e. it vanishes when the intermediate state is an on-shell tachyon.

This is general. Poles appear in the scattering amplitude when intermediate states go on-shell. The coefficients are precisely the couplings of the external states to the (nearly) on-shell physical state; this follows from the OPE.

21.6 The S-matrix vs. the effective action

The Virasoro–Shapiro and Veneziano amplitudes are beautiful formulas. Analogous formulas for the case of massless particles can be obtained. These are particularly important for the superstring. For many of the questions which interest us, we are not directly interested in the S-matrix. One feature of the string S-matrix construction is that it involves on-shell states; the momenta appearing in the exponential factors satisfy $p^2 = -m^2$, where m is the mass of the state. So one cannot calculate, for example, the effective potential for the tachyon, since this requires that all momenta vanish. For massless particles things are better, since $p = 0$ is the limiting case of an on-shell process. But the S-matrix is not precisely the effective action. Instead, given the S-matrix, it is usually a straightforward matter to determine a low-energy effective action which will reproduce it. At tree level, one just needs to subtract massless particle exchanges. In loops, one must be more careful.

It is particularly easy to extract three-point couplings of massless particles at tree level. One just needs to study an "S-matrix" for three particles (one can be a bit more careful and study a four-particle amplitude, isolating the coefficient of the massless propagator). From our previous analysis, we need

$$A = \frac{1}{\Omega_M} \langle V_1(z_1) V_2(z_2) V_3(z_3) \rangle, \tag{21.120}$$

where we don't integrate over the locations of the vertex operators. We are free to take z_1 and z_2 arbitrarily close to one another. Then the operator product will involve

$$V_1(z_1) V_2(z_2) \approx C_{123} \frac{1}{|z_1 - z_2|^2} V_3(z_2). \tag{21.121}$$

The final correlation function follows from the normalization of the vertex operators and cancels the Möbius volume. So the net result is that $g_s C_{123}$ is the coupling.

As an example, consider the coupling of two gravitons in the bosonic string. The vertex operator is

$$V_1 = \epsilon_{\mu\nu}(k_1) \partial X^\mu(z) \bar{\partial} X^\nu(z) e^{ik_1 \cdot X(z)}, \tag{21.122}$$

and similarly for V_2 and V_3. So the operator product has the structure:

$$V_1(z)V_2(w) = \frac{1}{|z-w|^4} + \epsilon_{\mu\nu}(k_1)\epsilon_{\rho\sigma}(k_2)e^{i(k_1+k_2)\cdot X(z)}\left(k_1^\nu k_2^\sigma \frac{1}{|z-w|^2}\partial X^\mu(z)\bar{\partial}X^\rho(z) + \cdots\right). \quad (21.123)$$

Here the first term arises from the contraction of all of the ∂X terms with each other. Loosely speaking, it is related to the production of off-shell tachyons. We will ignore it. The second term that we have indicated explicitly comes from contracting the first $\bar{\partial}X$ factor with the second exponential, and the second ∂X factor with the first exponential. The dots indicate a long set of contractions. The complete vertex is precisely the on-shell coupling of three gravitons in Einstein's theory, along with couplings to the antisymmetric tensor and dilaton. We will not worry with the details here. When we discuss the heterotic string, we will show that the theory completely reproduces the Yang–Mills vertex in much the same way. We shouldn't be surprised that it is difficult to define off-shell Green functions. In gravity, apart from the S-matrix, it is hard to define generally coordinate-invariant observables.

21.7 Loop amplitudes

So far, we have considered tree amplitudes. Closed or open strings interact by splitting and joining. Once we allow for quantum fluctuations, strings in intermediate states can split and join. Because of conformal invariance, the only invariant characteristic of these diagrams is their topology (for closed strings, the tree level world sheet has the topology of a sphere). In the closed string case, each additional loop adds a handle to the world sheet. In general, the theory of string loops is complicated. But the description of one-loop diagrams is rather simple, and exposes important features of the theory not apparent in tree diagrams. In the case of closed strings, requiring that the one-loop amplitude be sensible places strong constraints on the theory. Invariance under certain (global) two-dimensional general coordinate transformations, known as modular transformations, will account for many of the features of both the bosonic and superstring theories. In space-time, satisfying these constraints is a necessary condition for the unitarity of the scattering amplitude. In this section, we provide only a brief introduction. We will leave for later the discussion of open string loops.

The one-loop amplitude has the topology of a donut, or torus. A simple representation of a torus is as indicated in Fig. 21.2. In this figure, the world sheet is flat and of finite size. We can think of this torus as living in the complex plane. It is (up to conformal transformations) the world sheet appearing in the Euclidean path integral. The two possible periods of the donut are translated into two complex

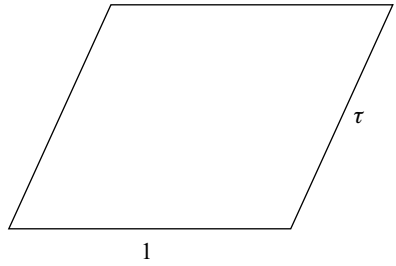

Fig. 21.2. A simple representation of a torus.

periods, λ_1 and λ_2. We require that the fields are periodic under

$$z \approx z + m\lambda_1 + n\lambda_2. \tag{21.124}$$

We can transform λ_1 and λ_2 by a transformation in the "modular group," $SL(2, Z)$:

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \lambda_1' \\ \lambda_2' \end{pmatrix} \tag{21.125}$$

with a, b, c and d integers satisfying $ad - bc = 1$, provided we also transform the integers n and m by the inverse matrix:

$$\begin{pmatrix} m \\ n \end{pmatrix} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \begin{pmatrix} m' \\ n' \end{pmatrix} \tag{21.126}$$

Now rescale z by λ_1, and call $\tau = \lambda_2/\lambda_1$. Then z has the periodicities 1 and τ. Under modular transformations, τ transforms as:

$$\tau \to \frac{a\tau + b}{c\tau + d}. \tag{21.127}$$

The modular transformations are general coordinate transformations of the world sheet theory, but they are not continuously connected to the identity. In order that one-loop string amplitudes make sense, we require that they are invariant under this transformation. The general amplitude will be a correlation function

$$\langle V(z_1)V(z_2)\ldots\rangle_{\text{torus}} \tag{21.128}$$

evaluated on the torus. The simplest amplitude is that with no vertex operators inserted. (At tree level, this amplitude vanishes owing to the division by the infinite Möbius volume.) For the bosonic string, we can evaluate the amplitude in light cone gauge. We simply need to evaluate the functional determinant. As these are free fields on a flat space, this is not too difficult. It is helpful to remember some basic field theory facts. The path integral, with initial configuration $\phi_i(x)$ and final

configuration $\phi_{\rm f}(x)$ computes the quantum mechanical matrix element:

$$\langle\phi_{\rm f}|e^{-iHT}|\phi_{\rm i}\rangle. \tag{21.129}$$

If we take the time to be Euclidean, impose periodic boundary conditions, and sum (integrate) over all possible $\phi_{\rm i}$, we have computed:

$$\mathrm{Tr}\; e^{-HT} \tag{21.130}$$

i.e. the quantum mechanical partition function. As described in Appendix C, this observation is the basis of the standard treatments of finite temperature phenomena in quantum field theory. In the present case, the periodicity is in the τ direction. So we compute

$$\mathrm{Tr}\; e^{-H_{\ell.c.}\tau}. \tag{21.131}$$

It is convenient to rewrite the light cone Hamiltonian, $H_{\ell.c,}$ in terms of L_0 and $\bar{L}_0$. Introducing

$$q = e^{2\pi i\tau} \quad \bar{q} = e^{-2\pi i\bar{\tau}} \tag{21.132}$$

we want to evaluate:

$$\mathrm{Tr}\; q^{L_0}\,\bar{q}^{\bar{L}_0}. \tag{21.133}$$

From any one oscillator with oscillator number n, just as in quantum mechanics, we obtain $(1-q^n)^{-1}$; so allowing for the different values of n and the $D-2$ transverse directions, we have:

$$\prod q^{D/24}\bar{q}^{D/24}(1-q^n)^{2-D}(1-\bar{q}^n)^{2-D}. \tag{21.134}$$

This is conveniently expressed in terms of a standard function, the Dedekind η function:

$$\eta(q) = q^{1/24}\prod_{n=1}^{\infty}(1-q^n). \tag{21.135}$$

We also need the contribution of the zero modes. This is:

$$\int \frac{d^{D-2}p}{(2\pi)^{D-2}} e^{-\tau_2 p^2} \propto \tau_2^{D-2}. \tag{21.136}$$

In the final expression, we need to integrate over τ. The measure for this can be derived from the Faddeev–Popov ghost procedure, but it can be guessed from the requirement of modular invariance. It is easy to check that

$$\int \frac{d^2\tau}{\tau_2^2} \tag{21.137}$$

is invariant. So, in 26 dimensions, we finally have

$$Z \propto \int \frac{d^2\tau}{\tau_2^2} \tau_2^{-12} |\eta(\tau)|^{-48}. \tag{21.138}$$

Now to check that this is modular invariant, we note, first, that the full modular group is generated by the transformations:

$$\tau \to \tau + 1 \quad \tau \to -1/\tau. \tag{21.139}$$

Under these transformations, as we said, the measure is invariant. The η function transforms as:

$$\eta(\tau + 1) = e^{\frac{i\pi}{12}} \eta(\tau) \quad \eta(-1/\tau) = (-i\tau)^{1/2} \eta(\tau) \tag{21.140}$$

Since $\tau_2 \to \tau_2/\tau_1^2 + \tau_2^2$, under $\tau \to -1/\tau$, Z is invariant. Here we see that the bosonic string makes sense only in 26 dimensions.

Suggested reading

More detail on the material in this chapter can be found in Green *et al.* (1987) and in Polchinski (1998). The light cone treatment described here is nicely developed in Peskin (1985).

Exercises

(1) Enumerate the states of the bosonic closed string at the first level with positive mass-squared. Don't worry about organizing them into irreducible representations, but list their spins.

(2) OPEs: explain why X^μ and X^ν do not have a sensible operator product expansion. Work out the OPE of ∂X^μ and ∂X^ν as in the text. Verify the commutator of α^μ and α^ν as in the text.

(3) Work out the Virasoro algebra, starting with the operator product expansion for the stress tensor, and using the contour method.

(4) Mermin–Wagner–Coleman Theorem: consider a free two-dimensional quantum field theory with a single, massless, complex field, ϕ. Describe the conserved $U(1)$ symmetry. Show that correlation functions of the form

$$\left\langle e^{iq_1\phi(x_1)} \dots e^{iq_n\phi(x_n)} \right\rangle \tag{21.141}$$

are non-vanishing only if $\sum q_i = 0$. Argue that this means that the global symmetry is not broken. From this construct an argument that global symmetries are never broken in two dimensions.

(5) Show that Ω_M of Eq. (21.109) is invariant under the Möbius group. You might want to proceed by analogy to the Faddeev–Popov procedure in gauge theories.

(6) Show that the factorization of tree-level S-matrix elements is general, i.e. that if the kinematics are correctly chosen for two incoming particles, 1 and 2, so that $(p_1 + p_2)^2 \approx m_n^2$, so that the amplitude is approximately a product of the coupling of particles 1 and 2 to n, times a nearly on-shell propagator for the field n.

22

The superstring

The theories we have described were motivated by thinking of a picture of a string moving in space-time. We arrived in this way at a description of strings in terms of two-dimensional quantum fields. The theories, so far, are theories of bosons only. But in this more abstract picture, we can imagine adding two-dimensional fermionic fields as well. This possibility was first considered by Ramond, Neveu and Schwarz, and leads to the superstring theories: Type I, Type IIA and Type IIB, and the two heterotic string theories. We first develop the theories in light cone gauge, where their spectra are readily exhibited. Then we discuss interactions.

22.1 Open superstrings

A priori, there appears to be a great deal of freedom in how we introduce fermions: their number, their representations under the (space-time) Lorentz group, and possibly other options. Various consistency conditions restrict these choices. In the case of open strings, we have to introduce one fermion, ψ^I, for each coordinate, X^I. For the action of the fermions we take:

$$S_\psi = \frac{1}{2\pi}\int d^2\sigma\, i\bar{\psi}^I \partial_\alpha \gamma^\alpha \psi^I. \tag{22.1}$$

In two dimensions, a particularly simple choice for the γ-matrices is:

$$\gamma^0 = \sigma_2 \quad \gamma^1 = i\sigma_1 \tag{22.2}$$

and the analog of γ_5 in four dimensions is

$$\gamma_3 = \sigma_3. \tag{22.3}$$

The Dirac equation, in this basis, is purely imaginary, so we can take the fermions to be real (Majorana). We can work with eigenfunctions of σ_3:

$$\psi^I = \begin{pmatrix} \psi^I_- \\ \psi^I_+ \end{pmatrix}. \tag{22.4}$$

In this way, if we again introduce light cone coordinates on the world sheet,

$$\sigma^{\pm} = \tau \pm \sigma \tag{22.5}$$

the action becomes:

$$S_{\psi} = \frac{1}{2\pi} \int d^2\sigma (\psi_+^I \partial_- \psi_+^I + \psi_-^I \partial_+ \psi_-^I). \tag{22.6}$$

We need to impose boundary conditions at the string end points. To determine suitable boundary conditions, we vary the Lagrangian to obtain the Euler–Lagrange equations. The surface terms which arise in the variation involve $\psi_+ \delta\psi_+ - \psi_- \delta\psi_-$. So the boundary terms vanish if $\psi_+ = \pm\psi_-$. An overall sign doesn't matter, so we can take the + sign at $\sigma = 0$:

$$\psi_+^I(0, \tau) = \psi_-^I(0, \tau) \tag{22.7}$$

This leaves two choices for the boundary conditions at $\sigma = \pi$:

$$\psi_+^I(\pi, \tau) = \pm\psi_-^I(\pi, \tau). \tag{22.8}$$

Fermions which obey the boundary condition with the + sign are called Ramond fermions; those with the − sign are called Neveu–Schwarz (NS) fermions. Corresponding to the Ramond case are the mode expansions:

$$\psi_-^I = \frac{1}{\sqrt{2}} \sum_{n \in Z} d_n^I e^{-in(\tau - \sigma)} \quad \psi_-^I = \frac{1}{\sqrt{2}} \sum_{n \in Z} d_n^I e^{-in(\tau + \sigma)}. \tag{22.9}$$

In the NS case we have:

$$\psi_-^I = \frac{1}{\sqrt{2}} \sum_{r \in Z + \frac{1}{2}} b_r^I e^{-ir(\tau - \sigma)} \quad \psi_-^I = \frac{1}{\sqrt{2}} \sum_{r \in Z + \frac{1}{2}} b_r^I e^{-ir(\tau - \sigma)}. \tag{22.10}$$

Now we quantize these fields:

$$\{\psi^I(\sigma, \tau)_{\pm}, \psi^J(\sigma', \tau)_{\pm}\} = \pi\delta(\sigma - \sigma')\delta^{IJ}\delta_{\pm\pm} \tag{22.11}$$

This gives, for the modes:

$$\left\{b_r^I, b_s^J\right\} = \delta^{IJ}\delta_{r+s} \quad \left\{d_m^I, d_n^J\right\} = \delta^{IJ}\delta_{m+n} \tag{22.12}$$

The Hamiltonian in light cone gauge, for the Ramond sector, is:

$$H = \vec{p}^2 + N_{\alpha} + N_d. \tag{22.13}$$

Here the Ns are the various number operators:

$$N_{\alpha} = \sum_{m=1}^{\infty} \alpha_{-m} \cdot \alpha_m \quad N_d = \sum_{m=1}^{\infty} m d_{-m} d_m. \tag{22.14}$$

For the NS sector, N_d is replaced by N_b:

$$N_b = \sum_{r=1/2}^{\infty} m b_{-r} b_r. \tag{22.15}$$

Each of these Hamiltonians has a normal ordering constant. We will determine these shortly. The states of the theory are the eigenstates of the fermion number operators $b_n^\dagger b_n$, $d_n^\dagger d_n$, etc., for non-zero n. The eigenvalues can take the values 0 or 1 in each case. The zero modes, which arise in the Ramond sector, are special. They give rise to space-time fermions.

22.2 Quantization in the Ramond sector: the appearance of space-time fermions

Usually, we do field theory at infinite volume, but here we are considering field theory at a finite volume ($0 < \sigma < \pi$), and this has introduced some new features. For the bosonic fields, X^I, we have already seen that there are zero modes, which gave rise to the coordinates and momenta of space-time. For the fermions, we now have the new feature that there are two sectors, with two independent Hilbert spaces. It is tempting to simply keep one, but it turns out that when we consider string interactions, it is necessary to include both: even if we attempt to exclude, say, the Ramond states, they will appear in string loop diagrams.

There is another feature: the appearance of fermion zero modes (d_0^I) in the Ramond sector. These are not conventional creation and annihilation operators. They obey the commutation relations:

$$\left\{d_0^I, d_0^J\right\} = \delta^{IJ}. \tag{22.16}$$

These are, up to a factor of 2, the anticommutation relations of Dirac gamma matrices for a $D-2$-dimensional space, i.e. they are associated with the group $O(D-2)$. Anticipating the fact that $D = 10$, we are interested in the Dirac matrices of $O(8)$. Before giving a construction of the spinor representations of $O(8)$, let us first simply state the basic result: $O(8)$ has two spinor representations, 8_s and $8'_\mathrm{s}$, and a vector representation, 8_v, all 8-dimensional. So we can realize the commutation relations, not on a Fock space, but on one of the 8-dimensional representations of $O(8)$. Labeling these states a, $\dot{a}$, then

$$\left\langle \dot{a} \middle| d_0^I \middle| a \right\rangle = \frac{1}{\sqrt{2}} \gamma^I_{\dot{a}a}. \tag{22.17}$$

We can construct an explicit representation for these matrices in various ways. A simple, and easy to remember construction, is to think of $O(8)$ as acting on eight

coordinates, x^I. Group these into complex coordinates:

$$z^1 = x^1 + ix^2 \quad z^2 = x^3 + ix^4 \quad z^3 = x^5 + ix^6 \quad z^4 = x^7 + ix^8 \tag{22.18}$$

and their complex conjugates. This defines an embedding of $U(4)$ in O(8). Correspondingly, we define

$$a^1 = (d_0^1 + id_0^2), \tag{22.19}$$

etc. The a^is obey the commutation relations:

$$\{a^i, a^{j\dagger}\} = \delta^{ij}, \tag{22.20}$$

all others vanishing. These are just the conventional anticommutation relations of fermion creation and annihilation operators (but remember for this discussion, these are just matrices, and shouldn't be confused with the d_ns, which are genuinely creation and annihilation operators). Among products of these operators we can distinguish two classes: those built from an even number of as, and those built from an odd number. In four dimensions, the analogous distinction corresponds to the eigenvalue (± 1) of γ^5.

Now we define a state, $|0\rangle$, annihilated by the a^is. We can then form two sets of states, those with "even fermion number" and those with odd. The even states are:

$$|0\rangle \quad a^{i\dagger}a^{j\dagger}|0\rangle \quad a^{1\dagger}a^{2\dagger}a^{3\dagger}a^{4\dagger}|0\rangle. \tag{22.21}$$

These states form one of the 8 representations, say 8_s. The second is formed by the states of odd fermion number. States are now labeled $|p^I, a, \{\text{oscillators}\}\rangle$.

What we have learned is that the states in the Ramond sector are space-time *fermions*; the states in the NS sector are space-time bosons.

22.3 Type II theory

For closed strings, we still have two-component fields ψ, but the possible choices of boundary conditions are somewhat different. We still require that the fermion surface terms vanish, but we also require that currents such as $\psi_+^I \psi_+^J$ be periodic. (These currents are part of the generators of rotations in space-time.) So we impose Ramond and Neveu–Schwarz boundary conditions independently on the left and right movers. Indeed, we treat the left- and right-moving fermions as independent fields. Recalling that the Lagrangian for the fermions breaks up into left- and right-moving parts. They have the mode expansions:

$$\psi^I = \sum_{n \epsilon Z} d_n^I e^{-2in(\tau-\sigma)} \quad \psi^I = \sum_{n \epsilon Z+1/2} b_r^I e^{-2ir(\tau-\sigma)} \tag{22.22}$$

in the Ramond and NS sectors, respectively, and

$$\tilde{\psi}^I = \sum \tilde{d}_n^I e^{-2in(\tau+\sigma)} \tilde{\psi}^I = \sum \tilde{b}_r^I e^{-2ir(\tau+\sigma)}. \tag{22.23}$$

The light cone Hamiltonian is now:

$$H = p^2 + N_\alpha + \tilde{N}_\alpha + N_d + \tilde{N}_d - a. \tag{22.24}$$

In constructing the spectrum, this must be supplemented with the condition of invariance under shifts in σ; in the covariant formulation, this was the $L_0 = \tilde{L}_0$ constraint.

22.4 World sheet supersymmetry

Before considering the spectrum, we consider the question of supersymmetry. The theory we are considering is supersymmetric in two dimensions. Just as we decomposed the fermions into left and right movers, we can introduce a two-component anticommuting parameter θ:

$$\theta = \begin{pmatrix} \theta_- \\ \theta_+ \end{pmatrix}. \tag{22.25}$$

Then we define the superfield:

$$Y^I = X^I + \bar{\theta}\psi^I + \frac{1}{2}\bar{\theta}\theta B^I. \tag{22.26}$$

We will see shortly that B^I is an auxiliary field, which in the case of strings in flat space, we can set to zero by its equations of motion. The supersymmetry generators are:

$$Q_A = \frac{\partial}{\partial\bar{\theta}_A} + i(\gamma^\alpha\theta)_A\partial_\alpha \tag{22.27}$$

(we are using the capital letter A for two-dimensional spinor indices here, to distinguish them from the letter a, which we used for $O(8)$ spinor indices, and the letter α, which we used for two-dimensional vector indices). As in four dimensions, we can introduce a covariant derivative operator which anticommutes with the supersymmetry generators:

$$D = \frac{\partial}{\partial\bar{\theta}} - i\gamma^\alpha\theta\partial_\alpha \tag{22.28}$$

In terms of the superfields, the action may be written in a manifestly invariant way:

$$\begin{aligned} S &= \frac{i}{4\pi}\int d^2\sigma d^2\theta \bar{D}Y^\mu DY_\mu \\ &= \frac{-1}{2\pi}\int d^2\sigma(\partial_\alpha X^I\partial^\alpha X^I - i\bar{\psi}^I\gamma^\alpha\partial_\alpha\psi^I_\mu - B^I B^I). \end{aligned} \tag{22.29}$$

Note that B^I vanishes by its equations of motion.

Finally, note that in the NS sectors, the boundary conditions explicitly break the world sheet supersymmetry; they map bosonic fields into fermionic fields,

and vice versa, which obey different boundary conditions. The Ramond sector is supersymmetric.

In the covariant formulation, this supersymmetry is essential to understanding the full set of constraints on the states. But it is important to stress that it is a symmetry of the world sheet theory; its implications for the theory in space-time are subtle.

22.5 The spectra of the superstrings

We have, so far, considered the world sheet structure of the superstring theories. We have not yet explored their spectra in detail. As in the case of the bosonic string, we will see that these theories possess a massless graviton. We will also find that they have a massless spin-3/2 particle, the gravitino. Consistent couplings of such a particle require that the space-time theory is supersymmetric.

22.5.1 The normal ordering constants

First, we give a general formula for the normal ordering constant. This is related to the algebra of the energy-momentum tensor we have discussed in Section 21.4. For a left- or right-moving boson, with modes which differ from an integer by η (e.g. modes are $1-\eta$, $2-\eta$, etc.), the contribution to the normal ordering constant is:

$$\Delta = -\frac{1}{24} + \frac{1}{4}\eta(1-\eta). \tag{22.30}$$

For fermions, the contribution is the opposite. So we can recover some familiar results. In the bosonic string, with 24 transverse degrees of freedom, we see that the normal ordering constant is -1. For the superstring, in the NS–NS sector, we have a contribution of $-1/24$ for each boson, and $1/24 - 1/16$ for each of the eight fermions on the left (and similarly on the right). So the normal ordering constant is $-1/2$. For the RR sector, the normal ordering vanishes.

There are simple derivations of this formula, whose justification requires careful consideration of conformal field theory. The normal ordering constant is just the vacuum energy of the corresponding two-dimensional free field theory. So we need

$$f(\eta) = \frac{1}{2}\sum_{1}^{\infty}(n+\eta). \tag{22.31}$$

Ignoring the fact that the sum is ill-defined, we can shift n by one, and compensate by a change in η:

$$f(\eta) = f(\eta+1) + \frac{1}{2}(1+\eta). \tag{22.32}$$

If we assume that the result is quadratic in η, we recover the formula above, up to the constant. We can "calculate" this constant by the following trick, known as zeta-function regularization. For $\eta = 0$, we need:

$$\sum_{n=1}^{\infty} n = \lim_{s \to -1} \sum_{n=1}^{\infty} n^{-s}. \tag{22.33}$$

The object on the right-hand side of this equation is $\zeta(s)$, the Riemann zeta function. The analytic structure of this function is something of great interest to mathematicians, but one well known fact is that its singularities lie off the real axis. Using integral representations, one can derive a standard result: $\zeta(-1) = -1/12$. This fixes the constant as $-1/24$. This argument may (should) appear questionable to the reader. The real justification comes from considering questions in conformal field theory.

22.5.2 The different sectors of the Type II theory

In the Type II theory, there are four possible choices of boundary conditions: NS for both left and right movers, Ramond for both left and right movers, Ramond for left and NS for right and NS for left and R for right. We will refer to these as the NS–NS, R–R, R–NS and NS–R sectors. Consider, first, the NS–NS sector. There are no zero-mode fermions, so we just have a normal (unique) ground state for the oscillators. From our computation of the normal ordering constants in the previous section, we see that $a = -1/2$ for both left and right movers. The lowest state is simply the state $|\vec{p}\rangle$. It has mass-squared -1 (in units with $\alpha' = 2$). Since no oscillators are excited, the $L_0 = \tilde{L}_0$ condition is satisfied. Now consider the first excited states. Again, we must have invariance under σ translations, so these are the states:

$$\psi^I_{-1/2} \tilde{\psi}^J_{-1/2} |\vec{p}\rangle. \tag{22.34}$$

Because $a = -1/2$ for both left and right movers, these states are massless. The symmetric combination here contains a scalar and a massless spin-two particle, the graviton; the antisymmetric combination is an antisymmetric tensor field. At the next level, we can create massive states using four space-time fermions or two bosons, or one fermion or two bosons.

Let's turn to the other sectors. Consider, first, the R–NS sector, where ψ is Ramond, $\tilde{\psi}$ is NS. Now, the left-moving normal ordering constant is zero, while the right-moving constant is $-1/2$. So we can satisfy the level-matching condition (invariance under σ translations) if we take the left movers to be in their ground state and take the right-moving NS state to be an excitation with a single fermion

operator above the ground state, i.e.:

$$\left|\Psi_a^I\right\rangle = \tilde{\psi}_{-1/2}^I |a\ \vec{p}\rangle. \tag{22.35}$$

From the space-time viewpoint, these are particles of spin-3/2 and 1/2. In the NS–R sector, we have another spin-3/2 particle.

Just as a massless spin-two particle requires that the underlying theory be generally covariant, a massless spin-3/2 particle, as we discussed in the context of four-dimensional field theories, requires *space-time* supersymmetry. But now we seem to have a paradox. With space-time supersymmetry, we can't have tachyons, yet our lowest state in the NS–NS sector, $|\vec{p}\rangle$, is a tachyon.

The solution to this paradox was discovered by Gliozzi, Scherk and Olive, who argued that it is necessary to project out states, i.e. to keep only states in the spectrum which satisfy a particular condition. This projection, which yields a consistent supersymmetric theory, is known as the GSO projection. Note, first, that we have been a bit sloppy with the fermion indices on the ground states. We have two types of fermion indices, a and $\dot{a}$, corresponding to the two spinor representations of $O(8)$. So we do the following. We keep only states on the left which are odd under left-moving world sheet fermion number; we do the same on the right, but we include in the definition of world sheet fermion number the chirality of the zero-mode states. We take

$$(-1)^F = e^{i\pi\gamma^9} \times e^{i\pi \sum_{1/2}^{\infty} \psi_n \psi_{-n}}. \tag{22.36}$$

In the R–NS sector, we make a similar set of projections. Here we have a choice, however, in which chirality we take. If we take the opposite chirality, we get the Type IIA theory; if we take the same chirality, we get the Type IIB theory.

Returning to the NS–NS sector, we make a similar projection, keeping only states which are odd under both left- and right-moving fermion number. In this way we eliminate the would-be tachyon in this sector.

Somewhat more puzzling is the R–R sector in each theory. Here both the left- and right-moving ground states are spinors. So in space-time, the states are bosons. We can organize them as tensors by constructing antisymmetric products of γ-matrices, $\gamma^{ijk\cdots}$. As we know from our experience in four dimensions, these form irreducible representations, in this case of the little group $O(8)$. Thinking of our construction of the γ-matrices in terms of the as, we can see γs with even numbers of indices connect states of opposite chirality, while those with odd numbers connect states with the same chirality. Which tensors appear depends on whether we consider the IIA or IIB theories. In the IIA case, only the tensors of even rank are non-vanishing. These tensors correspond to field strengths (one can consider an analogy with the magnetic moment coupling in electrodynamics, $\bar{\psi}\gamma^{\mu\nu}\psi$). So in the IIA theory, one has second- and fourth-rank tensors; the sixth- and eighth-rank field strengths are dual to these. In terms of gauge fields, there is a one-index tensor (a vector), and a

third-rank antisymmetric tensor. In the IIB theory, there is a scalar, a second-rank tensor and a fourth-rank tensor. In string perturbation theory, because the couplings are through the field strengths, there are no objects carrying the fundamental charge. Later we will see that there are non-perturbative objects, *D-branes*, which carry these charges.

22.5.3 Other possibilities: modular invariance and the GSO projection

The reader may feel that the choices of projections, and for that matter the choices of representations for the two-dimensional fermions, seem rather arbitrary. It turns out that the possible choices, at least for flat background space-times, are highly restricted. There are only a few consistent theories. Those we have described are the only ones without tachyons, and with both left- and right-moving supersymmetries on the world sheet.

In the bosonic string theory, we saw that it was crucial that the theory be formulated in 26 dimensions. One of the problems with the theory outside of 26 dimensions was that it is not modular invariant. This means that it is not invariant under certain global two-dimensional general coordinate transformations. This world sheet anomaly is correlated with anomalies in space-time. As for the gauge anomalies in field theories, these lead to breakdown of unitarity, Lorentz invariance, or both.

For the superstring theories, we will now explain why modular invariance demands a projection like the GSO projection. The point is that modular invariance relates sectors with different choices of boundary condition.

In our discussion of string theories up to this point, path integrals have appeared occasionally, but they are extremely useful in discussing string perturbation theory. The propagation of strings can be described by a two-dimensional path integral, with the string action, in much the same way as the amplitude for the motion of a particle. At tree level, the closed string world sheet has the topology of a sphere. At one loop, it has the topology of a torus. So at one loop, string amplitudes can be described as path integrals of a two-dimensional field theory on a torus. Note that we need, here, the full path integral, not simply the generator of Green's function for the field theory. The path integral on the torus, with no insertion of vertex operators, yields the partition function of the two-dimensional field theory. To understand this, let's consider the fermion partition function. Actually, there are several fermion partition functions. Let's begin with a single, right-moving Majorana fermion, and take, first, Neveu–Schwarz boundary conditions. There are two sorts of partition function we might define. First:

$$\mathrm{Tr}\, q^{L_0} = \prod_{r=1/2}^{\infty} (1 + q^r). \tag{22.37}$$

Alternatively, we can evaluate:

$$\mathrm{Tr}\,(-1)^F q^{L_0} = \prod_{r=1/2}^{\infty} (1 - q^r). \tag{22.38}$$

From a path integral point of view, the first expression is like a standard thermal partition function. It can be represented as a path integral with anti-periodic boundary conditions in the time direction. The second integral corresponds to a path integral with even boundary conditions for fermions in the time direction. We can represent the torus as in Fig. 21.2. Taking the vertical direction to be the time direction and the horizontal direction the space direction, we can indicate the boundary conditions with plus and minus signs along the sides of the square. Recalling the action of modular transformations on the torus, however, we see that the modular group mixes up the various boundary conditions. Not only does it mix the temporal boundary conditions, but it mixes the spatial boundary conditions as well.

It will be convenient for much of our later analysis to group the fermions in complex pairs. In the present case, this grouping is rather arbitrary, say $\Psi^1 = \psi^1 + i\psi^2$ and so on. Then the partition functions can be conveniently written in terms of θ functions. These functions, which have been extensively studied by mathematicians, transform nicely under modular transformations:

$$\vartheta \begin{bmatrix} \theta \\ \phi \end{bmatrix} (0, \tau) = \eta(\tau) e^{2\pi i \theta\phi} q^{\theta^2/2 - 1/24} \prod_{m=1}^{\infty} \left[1 + e^{2\pi i \phi} q^{m+\theta-1/2}\right] \times \left[1 + e^{-2\pi i \phi} q^{m-\theta-1/2}\right]. \tag{22.39}$$

Under $\tau \to \tau + 1$,

$$\vartheta \begin{bmatrix} \theta \\ \phi \end{bmatrix} (0, \tau + 1) = e^{i\theta^2 - \theta - \theta\phi} \vartheta \begin{bmatrix} \theta \\ \phi - \theta \end{bmatrix} (0, \tau) \tag{22.40}$$

while, under $\tau \to -1/\tau$,

$$\vartheta \begin{bmatrix} \theta \\ \phi \end{bmatrix} (0, 1/\tau) = e^{2\pi i \theta\phi} \vartheta \begin{bmatrix} -\phi \\ \theta \end{bmatrix} (0, \tau). \tag{22.41}$$

These transformation properties have a physical interpretation. Returning to Eqs. (21.125)–(21.127), the transformation $\tau \to -1/\tau$ exchanges the time and space directions of the torus. So these transformations interchange sectors with a given projection (multiplication of states by a phase) with states with a twist in the space direction. This is precisely what one would expect from a path integral, where boundary conditions in the time direction correspond to weighting of states with (symmetry) phases.

Calling

$$Z^{\alpha}_{\beta}(\tau) = \frac{1}{\eta(\tau)}\vartheta\begin{bmatrix}\alpha/2\\ \beta/2\end{bmatrix}(0,\tau) \tag{22.42}$$

the partition function for the eight fermions in the NS sector is $(Z^1_0)^4$, for example. If we include a $(-1)^F$ factor, this is replaced by $(Z^1_1)^4$. We can work out the similar expressions for the Ramond sector. From our expression for the transformation of the θ functions, it is clear that no one of these is modular invariant by itself, as we would expect from our path integral arguments. So it is necessary to combine them, and include also the eight bosons. When we do, we have the possibility of including minus signs (in more general situations, as we will see later, we will have more complicated possible phase choices). There are a finite number of possible choices. Two that work are:

$$Z_{\pm} = \frac{1}{2}\left[Z^0_0(\tau)^4 - Z^0_1(\tau)^4 + Z^1_0(\tau)^4 \mp Z^1_1(\tau)^4\right]. \tag{22.43}$$

These transform simply under the modular transformations; all of the terms transform to each other, up to an overall factor. There is a similar factor from the left-moving fermions (where one need not, a priori, take the same phase). Recall that the bosonic partition function is

$$Z_X(\tau) = (4\pi\alpha'\tau_2)^{-1/2}|\eta(q)|^{-2}. \tag{22.44}$$

Here the η function comes from the oscillators. The τ_2 factors come from the integration over the momenta. There are two additional such factors, coming from the integrals over the two light cone momenta. So the full partition function is:

$$Z = C\int \frac{d^2\tau}{\tau_2^2} Z_X^8 Z^+(\tau) Z^{\pm}(\tau)^*. \tag{22.45}$$

It is not hard to check that this expression is modular invariant.

If we examine the partition function carefully, we see that we have uncovered the GSO projection. Consider the first two terms in $Z^{\pm}$. This is just

$$\mathrm{Tr}(1-(-1)^F)_{\mathrm{NS}}, \tag{22.46}$$

i.e. it says that the physical states of the theory, in the NS sector, are only those of odd fermion number. There is a similar projector in the Ramond sector. The two possible choices of left- relative to right-moving Zs correspond precisely to the two possible supersymmetric string theories. Our original argument for the GSO projector was consistency in space-time, but here we have a more direct, world sheet consistency argument.

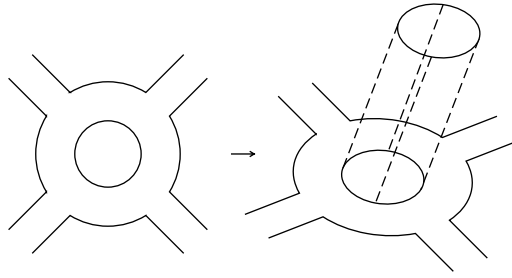

Fig. 22.1. Deforming the diagram for open string scattering reveals an intermediate closed string state.

These are the only choices of phases which lead to supersymmetric strings in ten dimensions. However, there are other choices which lead to non-supersymmetric strings. These give the Type 0 superstring. We will leave consideration of these theories to the exercises.

22.5.4 More on the Type I theory: gauge groups

In our discussion of the bosonic string theory, we mentioned that one can obtain non-Abelian gauge groups by allowing charges at the ends of the strings. There are an infinite set of possibilities, which we did not explore, as all of these theories have other problematic features if one is trying to describe Nature.

In the case of open superstrings, it turns out that the possible structures are quite constrained. First, it is necessary to include closed strings as well in order to obtain a unitary theory. This can be seen by considering scattering of four open strings. By stretching the diagram of Fig. 22.1, one can see that closed strings appear in intermediate states. These strings cannot be oriented. This leads to a different structure in the closed string sector than we saw in the IIA or IIB theories. It is necessary to require that states be symmetric under exchange of left- and right-moving quantum numbers. We will discuss the required projection later when we talk about D-branes and orientifold planes.

Second, it turns out that absence of anomalies fixes uniquely the gauge symmetry to be $O(32)$. From the point of view of our experience with four-dimensional anomalies, this is somewhat surprising, but it turns out that in ten dimensions supergravity by itself can be anomalous, and this is the case for the open string. Allowing for charges at the end of the string, leads to a set of additional mixed gauge and gravitational anomalies. Almost miraculously, if one takes the ends of the string to lie in the vector representation of $O(32)$, all anomalies cancel.

22.6 Manifest space-time supersymmetry: the Green–Schwarz formalism

In the Ramond–Neveu–Schwarz formalism, space-time supersymmetry is obscure. It only arises after imposing the GSO projector. The supersymmetry operators must connect the different sectors – essentially different two dimensional field theories. These operators can be constructed, though we will not do that in this text. Instead, we consider in this section a different formalism, the *Green–Schwarz formalism*, in which the space-time supersymmetry is manifest. This formalism is best understood in the light cone gauge.

In the Green–Schwarz formalism, one still has the bosonic coordinates, X^I, but the eight fermionic coordinates, ψ^I, in the vector representation of $O(8)$, are replaced by eight fermionic coordinates in one of the spinor representations of $O(8)$ (we have already seen that $O(8)$ possesses two spinor representations of opposite chirality). These are usually written as $S^a(\sigma, \tau)$. Their Lagrangian is:

$$\mathcal{L}_{\rm gs} = \frac{i}{2\pi}\bar{S}^a \rho^\alpha \partial_\alpha S^a, \tag{22.47}$$

where we have written the Ss as two component fermions, and ρ^α denotes the two-dimensional γ-matrices. The S_as can be taken real (Majorana). They can be decomposed into left and right movers, $S_\pm$. Unlike the case of RNS fermions, both for closed and open strings, one has only one boundary condition. As for the RNS fermions, for open strings, the boundary condition relates the left and right movers:

$$S^a_+(0, \tau) = S^a_-(0, \tau) \quad S^a_+(\pi, \tau) = S^a_-(\pi, \tau). \tag{22.48}$$

For the closed strings, one simply has periodicity,

$$S^a_\pm(\sigma + \pi, \tau) = S^a_\pm(\sigma, \tau). \tag{22.49}$$

The mode expansions, in the case of closed strings, are:

$$\begin{aligned} S^a_+ &= \sum_{-\infty}^{\infty} S^a_n e^{-2in(\tau-\sigma)} \\ S^a_- &= \sum_{-\infty}^{\infty} \tilde{S}^a_n e^{-2in(\tau+\sigma)}. \end{aligned} \tag{22.50}$$

The S_ns obey the anticommutation relations:

$$\left\{S^a_n, S^b_m\right\} = \delta^{ab}\delta_{m+n} \quad \left\{\tilde{S}^a_n, \tilde{S}^b_m\right\} = \delta^{ab}\delta_{m+n}. \tag{22.51}$$

For non-zero n these are canonical fermion creation and annihilation operator anticommutation relations. Because of their quantum numbers, the Ss, acting on space-time bosonic states, produce fermionic states, and vice versa.

The light cone Hamiltonian, in terms of these fields, takes the form:

$$H = \frac{1}{2p^+}((p^I)^2 + N + \tilde{N}), \tag{22.52}$$

where

$$N = \sum_{m=1}^{\infty} \left(\alpha^I_{-m}\alpha^I_m + m S^a_{-m} S^a_m\right) \quad \tilde{N} = \sum_{m=1}^{\infty} \left(\tilde{\alpha}^I_{-m}\tilde{\alpha}^I_m + m \tilde{S}^a_{-m} \tilde{S}^a_m\right). \tag{22.53}$$

Note that there is no normal ordering constant; more precisely, the normal ordering constants associated with the left- and right-moving fields vanish, because the contributions of the bosonic and fermionic fields cancel (as they do in the Ramond sector of the superstring).

As in the Ramond sectors of the superstring theories, the anticommutation relations of the zero modes are important and interesting:

$$\left\{S^a_0, S^b_0\right\} = \delta^{ab}. \tag{22.54}$$

Again they are similar to the anticommutation relations of Dirac γ-matrices, but now the indices are different than in the RNS case. The solution is to allow S_0 to act on sixteen states, eight of which carry spinor labels, $\dot{b}$, and eight of which carry $O(8)$ vector labels, I. Then

$$\left\langle I \middle| S^a_0 \middle| \dot{b} \right\rangle = \gamma^I_{a\dot{b}}. \tag{22.55}$$

We'll leave the verification of this relation for the exercises, and proceed directly to the identification of the massless states of the closed string theories. The IIA and IIB theories are distinguished by the relative helicities of the S and $\tilde{S}$ fields. In the IIA case, they are opposite; in the IIB case the same. The massless fields are just obtained by tensoring the left and right states of the zero modes. The states

$$\epsilon_{IJ}|I\rangle \times |J\rangle \tag{22.56}$$

are the graviton, B-field and dilaton; the states where $I \to a$ or $J \to a$, are the two gravitini of the theory; those where both I and J are replaced by spinor indices are the states we discovered in the Ramond–Ramond sector of the superstring theories.

In this formalism, the space-time supersymmetry is manifest. There are two types of supersymmetry generators. One generates not only space-time supersymmetries, but world sheet supersymmetries as well. This is as it should be; the world sheet Hamiltonian in the light cone gauge is also the space-time Hamiltonian.

$$Q^{\dot{a}} = \frac{1}{\sqrt{P^+}} \gamma^I_{a,\dot{a}} \sum_{-\infty}^{\infty} S^a_{-n} \alpha^I_n. \tag{22.57}$$

The second set are built of the zero modes alone:

$$Q^a = \sqrt{2P^+} S_0^a. \tag{22.58}$$

The supersymmetry generators obey the commutation relations:

$$\{Q^a, Q^b\} = 2P^+ \delta^{ab} \tag{22.59}$$
$$\{Q^a, Q^{\dot{b}}\} = \sqrt{2} \gamma^I_{a\dot{a}} P^I. \tag{22.60}$$
$$\{Q^{\dot{a}}, Q^{\dot{b}}\} = 2H \delta^{\dot{a}\dot{b}}. \tag{22.61}$$

The manifest supersymmetry and the close connection between world sheet and space-time supersymmetries makes the Green–Schwarz formalism a powerful tool, both conceptually and computationally, despite its lack of manifest Lorentz invariance.

22.7 Vertex operators

Because there are more world sheet fields in the superstring than in the bosonic string, the vertex operators are more complicated. In the RNS formalism, the supersymmetry on the world sheet is a relic of a larger, local supersymmetry, much as conformal invariance is a relic of the general coordinate invariance of the two-dimensional supersymmetry. The resulting superconformal symmetry provides constraints on vertex operators beyond those of the Virasoro algebra. These constraints can be implemented in a variety of ways, depending on how one treats the superconformal ghosts. In the simplest version, the vertex operators must be supersymmetric. In the case of the Type II theories, the vertex operators must respect both the left- and right-moving supersymmetries. For the massless fields of the Type II theory, for example:

$$V = \epsilon_{\mu\nu} (\bar{\partial} X^\mu - i k_\rho \psi^\rho \psi^\mu)(\bar{\partial} X^\nu - i k_\sigma \tilde{\psi}^\sigma \tilde{\psi}^\nu) e^{ik\cdot x}. \tag{22.62}$$

Here ϵ is subject to the constraint $k^\mu \epsilon_{\mu\nu} = 0$. Depending on the symmetries of ϵ, the vertex operator describes production of gravitons, dilatons, or antisymmetric tensor fields. It is straightforward to check that the coupling of three gravitons is that expected from the Einstein Lagrangian.

In the Green–Schwarz formalism, it is Lorentz invariance which governs the form of the vertex operators. As in the covariant formulation, the vertex operators in the Type II theory are products of separate vertex operators for the left and the right movers, with $e^{ik\cdot x}$ factors. These have the structure:

$$V_B = \zeta_{\mu\nu} B^\mu \tilde{B}^\nu e^{ik\cdot X}, \tag{22.63}$$

where

$$B^I = \partial X^I - R^{IJ}k^J \quad B^+ = p^+ \tag{22.64}$$

and, from the light cone gauge condition, $\zeta^{\mu+} = 0$. Here,

$$R^{IJ} = \frac{1}{4}\gamma^{IJ}_{ab} S^a S^b. \tag{22.65}$$

In the Green–Schwarz approach, it is not more difficult to deal with vertex operators for fermions or for what we have, in the covariant formulation, called the R–R states. The polarizations, $\zeta_{\mu\nu}$, are replaced by polarizations with one or two spinor indices. Then, as appropriate, one replaces the B^μs with fermionic operators, F^a and $F^{\dot{a}}$. We will not give these here as we will not need them in the text, but they can be found in the references. In the covariant approach, more conformal field theory machinery is required to construct fermion emission operators.

Suggested reading

The superstring is well treated in various textbooks. Green *et al.* (1987) focus heavily on the light cone formulation; Polchinski (1998) focuses on the RNS formulation. Both provide a great deal of additional detail, including construction of vertex operators and S-matrices in the two formalisms. A concise and quite readable introduction to the problem of fermion vertex operators in the RNS formulation is provided by the lectures of Peskin (1987).

Exercises

(1) Consider the R–R sectors of the IIA and IIB theories, and study the objects

$$\bar{u}\gamma^{IJK\cdots}u.$$

Show that in the IIA case, only even-rank tensors are non-vanishing, while in the IIB theory only the odd-rank tensors are non-vanishing. Phrase this in the language of ten dimensions, rather than the eight light cone dimensions. To do this consider a particle moving along the 9 direction, and show that the Dirac equation correlates chirality in ten dimensions with chirality in eight. To do this, you may want to make the following choice of Γ matrices:

$$\Gamma^0 = \sigma_2 \otimes I_{16};\ \Gamma^i = i\sigma_1 \otimes \gamma^i;\ \Gamma^9 = i\sigma_3 \otimes I_{16}. \tag{22.66}$$

(2) Write the Green–Schwarz Lagrangian in a superspace formulation. Show that $Q^{\dot{a}}$ is the supersymmetry generator expected in this approach. Construct the symmetry generated by Q^a, and show that this has the structure of a non-linearly realized (spontaneously broken) supersymmetry. Can you offer some interpretation?

(3) Verify that with the choice of Eq. (22.55), the zero modes of the Green–Schwarz operators S^a obey the correct anticommutation relations.

(4) Verify the expression for the partition function for the Type II theories. Show that it is modular invariant. Consider a different choice, which defines the type 0 superstring,

$$\left|Z_0^0\right|^8 + \left|Z_1^0\right|^8 + \left|Z_0^1\right|^8 \mp \left|Z_1^1\right|^8. \tag{22.67}$$

If you like, verify that this is also modular invariant, but at least show that the spectrum does not include a spin-3/2 particle.

(5) Verify that the operator product of two graviton vertex operators in the RNS formalism yields the correct on-shell coupling of three gravitons. Remember the gauge condition in this analysis. The three-graviton vertex in Einstein's theory can be found, for example, in Sannan (1986).

23

The heterotic string

In the Type II theory, we have seen that the left and right movers are essentially independent. At the level of the two-dimensional Lagrangian, there is a reflection symmetry between left and right movers. However, this symmetry does not hold sector by sector; it is broken by boundary conditions and projectors.

In the heterotic theory, this independence is taken further, and the degrees of freedom of the left and right movers are taken to be independent – and different. There are two convenient world sheet realizations of this theory, known as the fermionic and bosonic formulations. In both, there are eight left-moving and eight right-moving X^Is, associated with ten flat coordinates in space-time. There are eight right-moving two-dimensional fermions, ψ^I. There is a right-moving supersymmetry, but no left-moving supersymmetry. In the fermionic formulation there are, in addition, 32 left-moving fermions which have no obvious connection with space-time, λ^A. In the bosonic description, there are an additional 16 left-moving bosons. In other words, there are 24 left-moving bosonic degrees of freedom. There are actually several heterotic string theories in ten dimensions. Rather than attempt a systematic construction, we will describe the two supersymmetric examples. These have gauge group $O(32)$ and $E_8 \times E_8$. The group E_8, one of the exceptional groups in Cartan's classification, is not terribly familiar to most physicists. However, it is in this theory that we can most easily find solutions which resemble the Standard Model. We will introduce certain features of E_8 group theory as we need them. More detail can be found in the suggested reading. In this chapter, we will work principally in the fermionic formulation. We will develop some features of the bosonic formulation in later chapters, once we have introduced compactification of strings.

23.1 The $O(32)$ theory

The $O(32)(SO(32))$ theory is somewhat simpler to write down, so we develop it first. In this theory, the 32 λ^A fields are taken to be on an equal footing. The GSO projector, for the right movers, is as in the superstring theory. In the RNS formalism, in the NS sector, we keep only states of odd fermion number; similarly in the Ramond sector, where fermion number includes a factor $e^{i\Gamma_{11}}$. For the left movers, the conditions are different. Again, we have a Ramond and an NS sector. In the NS sector we keep states only of even fermion number. In the R sector, the ground state is a spinor of $SO(32)$. The spinor representation can be constructed just as we constructed the spinor representation of $O(8)$. Again, there are two inequivalent irreducible representations. There is a chirality, which we can call Γ_{33}. The lowest spinor representation of definite chirality is the 32 768. Again, in the Ramond sector, we project (by convention) onto states of even "fermion number."

As for the superstring, there is a different light cone Hamiltonian for each sector. The right-moving contributions are just as in the superstring. The left-moving part includes a contribution from the bosonic operators, and a contribution from the fermions, λ^A. As for the superstring, in the Ramond sector the λ^As are integer moded; they are half-integer moded in the NS sector. From our formula, the left-moving normal ordering constant is -1.

With this, we can consider the spectrum. Take, first, the NS–NS sector, i.e. the sector with NS boundary conditions for both the left and the right movers. The states are space-time bosons. The left-moving normal ordering constant is -1. Without λ^As, the lowest mass states we can form are:

$$\tilde{\alpha}^I_{-1}\psi^J_{-1/2}|0\rangle. \tag{23.1}$$

From our discussion of the normal ordering constants, we see that these states are massless. They have the quantum numbers of a graviton, antisymmetric tensor, and scalar field.

Using the left-moving fermion operators, we can construct additional massless states in this sector:

$$\lambda^A_{-1/2}\lambda^B_{-1/2}\psi^J_{-1/2}|0\rangle. \tag{23.2}$$

These are vectors in space-time. Because the λ^As are fermions, they are antisymmetric under $A \leftrightarrow B$. So they are naturally identified as gauge bosons of the gauge group $SO(32)$. We will show shortly that they have the couplings of $O(32)$ Yang–Mills theories.

Let's first consider the other sectors. In the NS–R sector, the right-moving states, $\psi^J_{-1/2}|\vec{p}\rangle$, are replaced by the states we labeled $|a\rangle$. Again, these must be massless, so we now have particles with the quantum numbers of the gravitino, one additional

fermion, and gauginos of $O(32)$. In the NS–R and R–R sectors, however, it turns out that there are no massless states, as can be seen by computing the normal ordering constants. It is necessary to include, as well, the R sector for the left movers. Here the normal ordering constant is $+1$, and there are no massless states.

23.2 The $E_8 \times E_8$ theory

The E_8 group is unfamiliar to many physicists, and one might wonder how one could obtain two such groups from a string theory. To begin, it is useful to note that E_8 has an $O(16)$ subgroup. Under this group, the adjoint of E_8, which is 248-dimensional, decomposes as a 120 – the adjoint of $O(16)$ – and a 128, a spinor of $O(16)$.

In ten dimensions, we have seen we can build a sensible string theory with eight left-moving bosons and 32 left-moving fermions. So the strategy is to break the fermions into two groups of 16, λ^A and $\lambda^{\tilde{A}}$, and to treat these as independent. This gives a manifest $O(16) \times O(16)$ symmetry, similar to the symmetry of the $O(32)$ theory. There are now NS and R sectors for each set of fermions separately. The right-moving GSO projectors are as before. For the left movers, in each of the NS sectors, the left-moving projector is onto states of even fermion number. With a suitable convention for the Γ_{11} chirality, this is also true of the R sectors. So consider, again, the spectrum. In the NS–NS–NS sector, just as before, there are a graviton, antisymmetric tensor, and scalar field. We can also construct gauge bosons in the adjoint of each of the two $O(16)$s:

$$\lambda^A \lambda^B \psi^J_{-1/2}|0\rangle \quad \lambda^{\tilde{A}} \lambda^{\tilde{B}} \psi^J_{-1/2}|0\rangle. \tag{23.3}$$

Note that because of the projectors, there are no massless states carrying quantum numbers of both $O(16)$ groups simultaneously. In the NS–NS–R sector, we find the superpartners of these fields.

Now consider the R–NS–NS sector. Here the ground state is a spinor of the first $O(16)$. So now we have a set of gauge bosons in the spinor 128-dimensional representation. Similarly, in the NS–R–NS sector, we have a spinor of the other $O(16)$. These are the correct set of states to form the *adjoints* of two E_8s. Again, establishing that the group is actually $E_8 \times E_8$ requires showing that the gauge bosons interact correctly. We will do that in the following section.

Finally, in the R–R–NS and R–R–R sectors, there are no massless states.

23.3 Heterotic string interactions

We would like to show that the states we have identified as gauge bosons in the heterotic string interact at low energies as required by Yang–Mills gauge invariance.

To do this, we work in the covariant formulation and construct vertex operators corresponding to the various states. Consider the $O(32)$ theory first. With our putative gauge bosons, we associate the vertex operators:

$$\int d^2z\ V^{AB\mu} = \int d^2z \lambda^A(\bar{z})\lambda^B(\bar{z})\left(\partial_z X^\mu(z) - ik_\nu \psi^\mu \psi^\nu(z)\right) e^{ik\cdot x}. \tag{23.4}$$

For the right movers, as in the Type II theories, we have required invariance under the right-moving world sheet supersymmetry. For the left-moving vertex operators, we have simply required that the operators have dimension one, so that overall the vertex operator has dimension one with respect to the left- and right-moving conformal symmetry (the operator is said to be (1, 1), just like those of the Type II theory). To determine their interactions, we study the operator product of two such operators. The left-moving part of the vertex operator is a current:

$$j^{AB}(\bar{z}) = \lambda^A(\bar{z})\lambda^B(\bar{z}). \tag{23.5}$$

The operator product of two of these currents is:

$$j^{AB}(\bar{z})j^{CD}(\bar{w}) = \frac{\delta^{AC}\delta^{BD} + \cdots}{(\bar{z}-\bar{w})^2} + \frac{\delta^{AC}\lambda^B(\bar{z})\lambda^D(\bar{w})}{(\bar{z}-\bar{w})}. \tag{23.6}$$

An algebra of currents of this kind is called a "Kac–Moody algebra." It has the general form

$$j^a(\bar{z})j^b(\bar{w}) = \frac{k\delta^{ab}}{(\bar{z}-\bar{w})^2} + \frac{f^{abc}j^c(\bar{w})}{(\bar{z}-\bar{w})}, \tag{23.7}$$

where k is called the central extension of the algebra. In our case, $k = 1$. The f^{abc}s are the structure constants of the group. This is what we have found here.

To see the Yang–Mills structure, it is helpful to use the general Kac–Moody form, denoting the currents, and the corresponding vertex operators, by a subscript a. In the operator product, we have seen from our discussion of factorization that the interaction is proportional to the coefficient of $1/|z-w|^2$. In the product $V_a(z)V_b(w)$, the $1/(\bar{z}-\bar{w})$ is proportional to f_{abc}, just what is needed for the Yang–Mills vertex. The momentum and $g_{\mu\nu}$ pieces arise from the right-moving operator product. In

$$(\partial X^\mu(z) + k_{1\rho}\psi^\rho(z)\psi^\mu(z))e^{ik_1\cdot X(z)}(\partial X^\nu(w) + k_{2\sigma}\psi^\sigma(w)\psi^\nu(w))e^{ik_2\cdot X(w)} \tag{23.8}$$

the $1/(z-w)$ terms arise from various sources. One can contract the ∂X factors in each vertex with the exponential factors. This gives

$$V_a^\mu V_b^\nu \sim \frac{f^{abc}V^{c\nu}\left(k_2^\mu - k_1^\mu\right)}{|z-w|^2}. \tag{23.9}$$

Contracting the two ∂X factors with each other gives two factors of $z - w$ in the denominator. These can be compensated by Taylor expanding $X(z)$ about w. Additional terms arise from contracting the fermions with each other. The details of collecting all the terms and comparing with the three gauge boson vertex are left for the exercises.

23.4 A non-supersymmetric heterotic string theory

One can verify the modular invariance of the heterotic string theory, with the GSO projections we have used, in precisely the same way as we did for the superstring theories. This raises the question: are there other ten-dimensional heterotic theories, obtained by combining the partition functions of the separate sectors in different ways? The answer is definitely yes. Several of these have tachyons, but one does not. Its gauge group is $O(16) \times O(16)$. It is most readily described in the Green–Schwarz formalism. This will also provide us with our first example of "modding out," obtaining a new string theory by making various projections.

In order to obtain the smaller gauge group, we need to get rid of the gauge bosons from E_8 which lie in the spinor representation. On the other hand, there is no harm in having the corresponding gauginos, if supersymmetry is broken. So we take the original $E_8 \times E_8$ theory, and keep only states which are even under the symmetry $(-1)^F$ in spacetime and a corresponding symmetry in the gauge group (i.e. spinorial representations are odd, non-spinorial even). This immediately gets rid of:

(1) the gravitinos, and
(2) the gauge bosons which are in spinorial representations of the group.

However, we have seen that, for consistency, it is important that string theories be modular invariant. Simply throwing away states spoils modular invariance; it is necessary to add in additional states. In the present case, one has to add a sector with different, twisted boundary conditions for the fields:

$$S_a(\sigma + \pi, \tau) = -S_a(\sigma, \tau). \tag{23.10}$$

For the gauge fermions there is a related boundary condition (this is more easily described in the bosonic formulation which we will discuss in the chapter on compactification).

Suggested reading

The original heterotic string papers by Gross *et al.* (1985, 1986) are remarkably clear. Polchinski's book (1998) provides a quite thorough overview of these theories.

For example, for those who are not enamored of the Green–Schwarz formalism, it develops the non-supersymmetric $O(32)$ in the RNS formalism in some detail.

Exercises

(1) Construct the states corresponding to the gauge bosons of $E_8 \times E_8$. In particular, use the creation–annihilation operator construction of $O(2N)$ spinor representations to build the 128s of $O(16)$.

(2) Verify that the algebra of $O(32)$ currents is of the Kac–Moody form. To work out the structure constants, remember that the generators of O groups are just the antisymmetric matrices:

$$(\omega^{AB})_{CD} = \delta^{AC}\delta^{BD} - \delta^{AD}\delta^{BC}. \tag{23.11}$$

(3) Verify that, on-shell, the three-gluon vertex has the correct form. In addition to carefully evaluating the terms in the operator product expansion, it may be necessary to use momentum conservation and the transversality of the polarization vectors.

24

Effective actions in ten dimensions

In ten dimensions, supersymmetry greatly restricts the allowed particle content and effective actions for theories with massless fields. Without gauge interactions, there are only two consistent possibilities. These correspond to the low-energy limits of the IIA and IIB theories. These have $N = 2$ supersymmetry (they have 32 conserved supercharges). Because the symmetry is so restrictive, we can understand a great deal about the low-energy limits of these theories without making any detailed computations. We can even make exact statements about the non-perturbative behavior of these theories. This is familiar from our studies of field theories in four dimensions with more than four supercharges. In ten dimensions, supersymmetric gauge theories have $N = 1$ supersymmetry (16 supercharges). Classically, specification of the gauge group completely specifies the terms in the effective action with up to two derivatives. Quantum mechanically, only the gauge groups $O(32)$ and $E_8 \times E_8$ are possible.

24.0.1 Eleven-dimensional supergravity

Rather than start with these ten-dimensional theories, it is instructive to start in eleven dimensions. Eleven is the highest dimension where one can write a supersymmetric action (in higher dimensions, spins higher than two are required). This fact by itself has focused much attention on this theory. But it is also known that eleven dimensions has a connection to string theory. As we will see later, if one takes the strong coupling limit of the Type IIA string theory, one obtains a theory whose low-energy limit is eleven-dimensional supergravity.

The particle content of the eleven-dimensional theory is simple: there is a graviton, g_{MN} (44 degrees of freedom) and a three-index antisymmetric tensor field, C_{MNO} (84 degrees of freedom). There is also a gravitino, ψ_M. This has (16×8) degrees of freedom. We have, as usual, counted degrees of freedom by considering a theory in nine dimensions, remembering that g_{MN} is symmetric and traceless,

and that the basic spinor representation in nine dimensions is sixteen dimensional (it combines the two eight-dimensional spinors of $O(8)$).

The Lagrangian for the eleven-dimensional theory, in addition to the Ricci scalar, involves a field strength for the three-index field, C_{MNO}. The corresponding field strength, F_{MNOP}, is completely antisymmetric in *its* indices, similar to the field strength of electrodynamics:

$$\begin{aligned} F_{MNOP} &= \frac{3!}{4!}\left(\partial_M C_{NOP} - \partial_N C_{MOP} + \cdots\right) \\ &= \frac{3!}{4!}\sum_P (-1)^P \partial_M C_{NOP}, \end{aligned} \tag{24.1}$$

where the sum is over all permutations, and the factor $(-1)^P$ is ± 1 depending on whether the permutation is even or odd. It is convenient to describe such antisymmetric tensor fields in the language of differential forms. For the reader unfamiliar with these, an introduction is provided later, in Section 26.1. For now, we note that antisymmetric tensors with p indices are p forms. The operator of taking the curl, as in Eq. (24.1), takes a p form to a $p+1$ form. It is denoted by the symbol d, and is called the *exterior derivative.* In terms of forms, Eq. (24.1) can be written quite compactly as

$$F = dC. \tag{24.2}$$

The theory has a gauge invariance:

$$C \to C + d\Lambda \quad C_{MNO} \to \frac{2}{3!}\sum_P (-1)^P \partial_M \Lambda_{NO} \tag{24.3}$$

where Λ is a two-form.

We will not need the complete form of the action. The bosonic terms are:

$$\mathcal{L}_{\text{bos}} = -\frac{1}{2\kappa^2}\sqrt{g}R - \frac{1}{48}\sqrt{g}F^2_{MNPQ} - \frac{\sqrt{2}\kappa}{3456}\epsilon^{M_1\ldots M_{11}} F_{M_1\ldots M_4} F_{M_5\ldots M_8} C_{M_9\cdots M_{11}}. \tag{24.4}$$

The last term is a Chern–Simons term. It respects the gauge invariance of Eq. (24.3) if one integrates by parts. Such terms can arise in field theories with odd dimensions; in 2+1-dimensional electrodynamics, for example, they play an interesting role. The fermionic terms include covariant derivative terms for the gravitino, as well as couplings to F and various four-Fermi terms. The supersymmetry transformation laws have the structure:

$$\delta e^A_M = \frac{\kappa}{2}\bar{\eta}\Gamma^A \psi_M \tag{24.5}$$

$$\delta A_{MNP} = -\frac{\sqrt{2}}{8}\bar{\eta}\Gamma_{[MN}\psi_{P]} \tag{24.6}$$

$$\delta\psi_M = \frac{1}{\kappa}D_M\eta + (F\eta \text{ pieces}). \tag{24.7}$$

Here e_M^A is the "vielbein" field, and the covariant derivative is constructed from the spin connection (discussed in Section 17.6).

24.0.2 The IIA and IIB supergravity theories

The eleven-dimensional fields are functions of the coordinates $x_0, \ldots, x_{10}$. We obtain the IIA supergravity theory (the low-energy limit of the Type IIA string) theory if we truncate the eleven-dimensional supergravity theory to ten dimensions, i.e. if we simply eliminate the dependence on x_{10}. We need to relabel fields, as well, since it is not appropriate to have a 10 index. So we take the components of g with ten-dimensional indices to be the ten-dimensional metric. Then $g_{10\,10}$ is a ten-dimensional scalar, which we call ϕ, and $g_{10\,\mu}$ is a ten-dimensional vector, which corresponds to the Ramond–Ramond vector of the IIA string theory. Note that $C_{11\,\mu\nu} = B_{\mu\nu}$ is a two-index antisymmetric tensor field in ten dimensions (corresponding to the two-index tensor we found in the NS–NS sector). The gravitino decomposes into two ten-dimensional gravitinos, and two spin-$1/2$ particles. With $H = dB$, the bosonic terms in the ten-dimensional action for the NS–NS fields are:

$$\mathcal{L}_{\text{bos}} = -\frac{1}{2\kappa^2}R - \frac{3}{4}\phi^{-3/2}H_{\mu\nu\rho}^2 - \frac{9}{16\kappa^2}(\partial_\mu\phi/\phi)^2. \tag{24.8}$$

The IIB theory is not obtained in this way. But from string theory, we can see that the NS–NS in the action must be the same as in the Type IIA theory. This is because in the NS–NS sector, the vertex operators of the IIA and IIB theories are the same, so the scattering amplitudes – and hence the effective action – are the same as well.

24.0.3 Ten-dimensional Yang–Mills theory

From our studies of the heterotic string, we know the field content of this theory. There is a metric, an antisymmetric tensor field (which we again call $B_{\mu\nu}$), a scalar ϕ, and the gauge fields, A_μ^a. The Lagrangian for g, B and ϕ is the same as in the Type II theories. The gauge terms are:

$$\mathcal{L}_{\text{YM}} = -\frac{\phi^{-3/4}}{4g^2}F_{\mu\nu}^2 - \frac{1}{2}\bar{\chi}^a(D_M\chi)^a. \tag{24.9}$$

It turns out that there is another crucial modification in the Yang–Mills case. The field strength H_{MNO} is not simply the curl of B_{MN} but contains an additional piece, which closely resembles the Chern–Simons term we encountered in our study of instantons in four-dimensional Yang–Mills theory:

$$H = dB - \frac{\kappa}{\sqrt{2}}\omega_3 \tag{24.10}$$

(the notation will be thoroughly explained in Chapter 26), with

$$\omega_3 = A^a F^a - \frac{1}{3} g f_{abc} A^a A^b A^c = A^a dA^a + \frac{2}{3} g f_{abc} A^a A^b A^c. \tag{24.11}$$

There is also a gravitational piece, with a similar form.

This extra term plays an important role in understanding anomaly cancellation. In four dimensions, we will see that it leads to the appearance of axions in the low-energy theory.

24.1 Coupling constants in string theory

The Standard Model is defined, in part, by specifying a set of coupling constants. The fact that there are so many parameters is one of the reasons we have given that the model is not satisfactory as some sort of ultimate description of nature. In our discussion of string interactions, we have introduced a coupling constant, g_s. There is one such constant for each of the string theories we have introduced: bosonic, Type I, Type IIA and Type IIB, and heterotic. But the idea that string theory possesses a free parameter is, it turns out, an illusion. By changing the expectation value of the dilaton field, we can change the value of the coupling. This is similar to phenomena we observed in four-dimensional supersymmetric gauge theories. In situations with a great deal of supersymmetry, there will be no potential, perturbatively or non-perturbatively, for this field, and the choice of coupling will correspond to a choice of vacuum. But in vacua in which supersymmetry is broken, we would expect that dynamical effects would fix the value of this and any other moduli. The coupling constants of the low-energy theory would then be determined fully in ways which, in principle, one could understand and eventually hope to calculate. In the next few sections, we explain this connection between coupling constants and fields.

24.1.1 Couplings in closed string theories

When we constructed vertex operators, we saw that we could include a coupling constant, g_s, in the definition of the vertex operator. In the heterotic string, the same coupling enters in all vertices. This is a consequence of unitarity. At tree level, for example, we saw that scattering amplitudes factorize near poles of the S-matrix. If one introduced independent couplings for each vertex operator, the amplitudes would not factorize correctly. As a result, all amplitudes can be expressed in terms of a single parameter. In the heterotic string theory, this means that there is a calculable relation between the gravitational constant and the Yang–Mills coupling. To work out this coupling, one needs to calculate the three-point interaction for three

gravitons, and for three gauge bosons carefully (see the exercises at the end of the chapter). The results are necessarily of the form:

$$\kappa_{10}^2 = a g^2 (2\alpha')^4 \quad g_{\rm YM}^2 = b g^2 (2\alpha')^3. \tag{24.12}$$

The calculation yields $a = 1/4$, $b = 1$.

A similar analysis, in the Type I theory, gives a relation between the open string and closed string couplings, and between the gauge and gravitational couplings.

In both theories, we see that the string scale is smaller than the Planck scale:

$$M_{\rm s} = (g_{\rm s})^{1/4} M_{\rm p}. \tag{24.13}$$

This is a satisfying result. It means that if we think of $M_{\rm s}$ as the cutoff on the gravity theory, gravitational loops are suppressed by powers of $g_{\rm s}$.

24.1.2 The coupling is not a parameter in string theory

So far, in all of the string theories, it appears that there is an adjustable, dimensionless parameter. As we said earlier, this is not really the case. The reason for this traces to the dilaton. Classically, in all of the string theories we have studied, the dilaton has no potential, so its expectation value is not fixed. In the next two short sections, we will demonstrate that changing the expectation value of the dilaton changes the effective coupling. With enough supersymmetry, there is no potential for the dilaton, so the question of the value of the coupling is equivalent to a choice among degenerate vacuum states. Without supersymmetry (or with $N \leq 1$ supersymmetry in four dimensions), one does expect quantum mechanical effects to generate a potential for the dilaton, and the value of the coupling is a *dynamical* question.

24.1.3 Effective Lagrangian argument

Perhaps the simplest way to understand the role of the dilaton is to examine the ten-dimensional effective action. Start with the case of the heterotic string in ten dimensions. We can redefine $\phi = g^{-2}\kappa^{3/2}\phi'$, eliminating g everywhere in the action. Note that, since $\kappa \propto g$, this means that $\phi' \sim g^{1/2}$. Then we can do a Weyl rescaling which puts a common power of ϕ in front of the action (dropping the prime on ϕ):

$$g_{\mu\nu} = \phi^{-1} g_{\mu\nu} \tag{24.14}$$

puts a common power of ϕ out front of the action: ϕ^{-4}. This is consistent with g being the string loop parameter, since we have effectively g^{-2} at the front.

With this rescaling, it is the string scale which is fundamental. Remember that $M_{\rm p}^2 = M_{\rm s}^2/(g^2)^{1/4}$. By rescaling the metric, we have rescaled lengths, which were originally expressed in units of $M_{\rm p}$, in terms of $M_{\rm s}$. So we have a consistent picture.

The cutoff for the effective Lagrangian is M_s. All dimensional parameters in the Lagrangian are of order M_s, and loops are accompanied by $g^2 \sim \phi^4$.

24.1.4 World sheet coupling of the dilaton

Just as we can couple the graviton to the world sheet, we can couple the dilaton. The dilaton turns out to couple to the two-dimensional curvature:

$$\mathcal{L}_\Phi = \frac{1}{4\pi} \int d^2\sigma \sqrt{h} \Phi(X) R^{(2)}. \tag{24.15}$$

In two dimensions, however, gravity is "trivial." If we use our usual counting rules, the graviton has -1 degree of freedom. So the $R^{(2)}$ term should not generate any sensible graviton dynamics. If we go to conformal gauge,

$$h_{\alpha\beta} = e^\phi \eta_{\alpha\beta} \tag{24.16}$$

the curvature is a total divergence:

$$R^{(2)} = \partial^2 \phi. \tag{24.17}$$

So at most this term in the action is topological. To get some feeling for this, let's evaluate the integral in the case of a sphere. We have seen that one representation for the sphere is provided by the space CP^1. This space has one complex coordinate. It is Kahler, which means that the only non-vanishing component of g is $g_{z\bar{z}}$:

$$g_{z\bar{z}} = (\partial_z \partial_{\bar{z}} K(z, \bar{z})) \tag{24.18}$$

where, in this case:

$$K = \ln(1 + \bar{z}z). \tag{24.19}$$

So

$$g = \left(\frac{1}{1 + \bar{z}z} \right)^2. \tag{24.20}$$

From this, we can read off ϕ:

$$\phi = 2\ln(1 + \bar{z}z) = -2\ln\left(1 + \sigma_x^2 + \sigma_y^2\right) \tag{24.21}$$

so the integral over the curvature is:

$$\frac{1}{4\pi} \int d^2\sigma \, \partial^2 \Big(-2\ln\left(1 + \sigma_x^2 + \sigma_y^2\right) \Big) = 2. \tag{24.22}$$

Note that this is invariant under a constant Weyl rescaling; it is topological. It is known as the Euler character of the surface, and satisfies:

$$\chi = \frac{1}{4\pi} \int d^2\sigma \sqrt{h} R^{(2)} \tag{24.23}$$

and

$$\chi = 2(1 - g). \tag{24.24}$$

In this expression, χ is known as the Euler character of the manifold, and g is the genus. For the sphere, $g = 0$; for the torus, $g = 1$, and so on for higher-genus string amplitudes. So string amplitudes, for constant Φ, come with a factor:

$$e^{-2\Phi(1-g)}. \tag{24.25}$$

So we can identify e^{Φ} with the string coupling constant.

Suggested reading

The ten-dimensional effective actions are described in some detail by Green *et al.* (1987). The couplings of the dilaton in string theory are discussed in detail by Polchinski (1998).

Exercise

(1) By studying the OPEs of the appropriate vertex operators, verify Eq. (24.12). To avoid making this calculation too involved, you may want to isolate particular terms in the gravitational and Yang–Mills couplings. The required vertices in general relativity can be found in Sannan (1986).

25

Compactification of string theory I. Tori and orbifolds

We don't live in a ten-dimensional world, and certainly not in a twenty-six-dimensional world without fermions. But if we don't insist on Lorentz invariance in all directions, there are other possible ways to construct consistent string theories. In this chapter we will uncover many consistent string theories in four dimensions (and in others). If anything, our problem will shortly be an embarrassment of riches: we will see that there are vast numbers of possible string constructions. The connection of these various constructions to one another is not always clear. Many of these can be obtained from one another by varying expectation values of light fields (moduli). One might imagine that others could be obtained by exciting massive fields as well. In general, though, this is not known, and, in any case, the meaning of such connections in a theory of gravity is obscure. But before exploring these deep and difficult questions, we need to acquire some experience with constructing strings in different dimensions.

25.1 Compactification in field theory: the Kaluza–Klein program

The idea that space-time might be more than four-dimensional was first put forward by Kaluza and Klein shortly after Einstein published his general theory of relativity. They argued that, in this case, *five-dimensional* general coordinate invariance would give rise to both four-dimensional general coordinate invariance and a $U(1)$ gauge invariance, unifying electromagnetism and gravity. In modern language, they considered the possibility that space-time is five-dimensional, with the structure $M^4 \times S^1$. This is, on first exposure, a bizarre concept, but its implications are readily understood by considering a toy model. Take a single scalar field, Φ, in five dimensions. Denote the coordinates of M^4 by x^μ, as usual, and that of the fifth dimension by y,

$$0 \leq y < 2\pi R. \tag{25.1}$$

Because y is a periodic variable, we can expand the field Φ in Fourier modes:

$$\Phi(x, y) = \sum_n \frac{1}{\sqrt{2\pi R}} \phi_n(x) e^{i p_n y} \quad p_n = \frac{n}{R}. \tag{25.2}$$

Taking a simple free field Lagrangian for Φ in five dimensions, the Lagrangian, written in terms of the Fourier modes, takes the form:

$$\begin{aligned} \int d^4x dy \mathcal{L} &= -\int d^4x dy \frac{1}{2} \left[(\partial\phi)^2 + M^2\phi^2) \right] \\ &= -\int d^4x \sum_n \frac{1}{2} \left(\partial_\mu \phi^2 + \left(M^2 + p_n^2 \right) \phi^2 \right). \end{aligned} \tag{25.3}$$

So, from a four-dimensional perspective, this theory describes an infinite number of fields, with ever increasing mass. In the gravitational case, symmetry considerations will force $M = 0$. If we set $M = 0$ in our scalar model, we obtain one massless state in four dimensions ($n = 0$), and an infinite tower – the Kaluza–Klein tower – of massive states. If R is very small, say $R \approx M_{\rm p}^{-1}$, the massive states are all extremely heavy. For the physics of the every-day world, we can integrate out these massive fields, and obtain an effective Lagrangian for the massless field. The effects of the infinite set of massive fields – the signature of extra dimensions – will show up in tiny, higher-dimension operators. So, in the end, finding evidence for these extra dimensions is likely to be extremely difficult.

Having understood this simple model, we can understand Kaluza and Klein's theory of gravitation and electromagnetism. The five-dimensional theory has the Lagrangian:

$$\mathcal{L} = \frac{1}{2\kappa^2} \sqrt{g} R. \tag{25.4}$$

Now there is an infinite tower of massive states, corresponding to modes of the five-dimensional metric: $g_{\mu\nu}$, $g_{\mu 4}$ and g_{44}. Our principal interest is in the massless states, which arise from modes which are independent of y (we will need to refine this identification shortly). We expect to find a four-dimensional metric tensor, $g_{\mu\nu}$, a field which transforms as a vector of the four-dimensional Lorentz group, $g_{4\mu}$, and a scalar: g_{44}. There are various ways we can rewrite the five-dimensional fields in terms of four-dimensional fields. The physics is independent of this choice, but clearly some choices will be better than others. The most sensitive choice is that of the gauge field; we would like to choose this field so that its gauge transformation properties are simple. The general coordinate invariance associated with transformations of the fifth dimension, $x_4 = x_4 + \epsilon_4(x)$, is:

$$g_{\mu 4} = g_{\mu 4} + \partial_\mu \epsilon_4(x). \tag{25.5}$$

This looks just like the transformation of a gauge field. So we adopt the conventions:

$$g_{\mu 4} = A_\mu;\ g_{44}(x) = e^{2\sigma(x)};\ g_{\mu\nu} = g_{\mu\nu}. \tag{25.6}$$

Note we are defining, here, a reference metric and measuring distances relative to that; we can take the basic distance to be the Planck length. Substituting this Ansatz back in the five-dimensional action, one can proceed very straightforwardly, working out the Christoffel symbols and from these the various components of the curvature. Gauge invariance significantly constrains the possible terms. One obtains:

$$\mathcal{L} = \frac{2\pi R}{2\kappa^2}\sqrt{g}e^{\sigma}(R) + \frac{1}{4}e^{-\sigma}F_{\mu\nu}^2. \tag{25.7}$$

So the theory, at low energies, consists of a $U(1)$ gauge field, the graviton, and a scalar. The Lagrangian is not quite in the canonical form; usually one writes the action for general relativity in a form where the coefficient of the Ricci scalar (the "Einstein term") is field-independent. One can achieve this by performing an overall rescaling of the metric, known as a Weyl rescaling,

$$g_{\mu\nu} \to e^{-\sigma} g_{\mu\nu}. \tag{25.8}$$

This introduces a kinetic term for the scalar:

$$\mathcal{L} = \frac{1}{2\kappa^2}(R + 3/2(\partial\phi)^2). \tag{25.9}$$

The scalar field here is particularly significant. As it corresponds to g_{55}, giving it an expectation value amounts to changing the radius of the internal space. In the Lagrangian, there is no potential for σ, so at this level, nothing determines this expectation value. As in our four-dimensional examples, σ is said to be a modulus. We now show that quantum mechanical effects generate a potential for σ, already at one loop. This potential falls to zero rapidly as the radius becomes large. If there is a minimum of the potential, it occurs at radii of order one, where the computation is certainly not reliable.

The calculation is equivalent to a Casimir energy computation in quantum field theory; one can think of the system as sitting in a periodic box of size $2\pi R$, and asking how the energy depends on the size of the box. We can guess the form of the answer before doing any calculation. Since this is a one-loop computation, the result is independent of the coupling. On dimensional grounds, the energy density is proportional to $1/R^4$.

To simplify matters, we will treat the gravitational field as a scalar field. At one loop:

$$\Gamma = \mathrm{Tr}\ln\left(-\partial^2 + \frac{n^2}{R^2}\right), \tag{25.10}$$

where we can do the calculation in Euclidean space. We can obtain a more manageable expression by differentiating with respect to R. The trace can be interpreted now as a sum over the possible momentum states in four Euclidean dimensions, in a box of volume VT. Replacing the sum by an integral gives an explicit factor of VT; the coefficient is the energy per unit volume:

$$\frac{\partial V}{\partial R} = \int \frac{d^4p}{(2\pi)^4 R^3} \sum_n \frac{n^2}{p^2 + (n^2/R^2)}. \tag{25.11}$$

This can be evaluated using the same trick one uses to compute the partition function in finite-temperature field theory (this is described in Appendix C). One first converts the sum into a contour integral, by introducing a function with simple poles located at the integers:

$$\frac{\partial V}{\partial R} = \int \frac{d^4p}{(2\pi)^3} \oint \frac{dz}{2\pi i} \frac{1}{z^2 + p^2} \frac{1}{1 - e^{2\pi i R z}} z^2/R. \tag{25.12}$$

The contour consists of one line running slightly above the real axis, and one line running below. Now deform the contour, so that the upper line encircles the pole at $z = ip$, and the lower line encircles the pole at $z = -ip$. The resulting expression is divergent, but we can separate off a piece independent of R and a convergent, R-dependent piece:

$$\begin{aligned} \frac{\partial V}{\partial R} &= \frac{1}{R} \int \frac{d^4p}{(2\pi)^4} \frac{p^2}{2p} \left[1 + \frac{1}{e^{2\pi p R} - 1} \right] \\ &= \frac{24\zeta[5]}{(2\pi)^4 R^5} + R\text{-independent}. \end{aligned} \tag{25.13}$$

25.1.1 Generalizations and limitations of the Kaluza–Klein program

So far we have considered compactification of a five-dimensional theory on a circle, but one can clearly consider compactifications of more dimensions on more complicated manifolds. It is possible to obtain, in this way, non-Abelian groups. So one might hope to understand the interactions of the Standard Model. The principal obstacle to such a program turns out to be obtaining chiral fermions in suitable representations. The existence of chiral fermions in a particular compactification is a topological question. As one varies the size and shape of the manifold, it is possible that some fields will become massless; equivalently, massless fields can become massive. But fields which gain mass must come in vector-like pairs. Chiral fermions will not simply appear or disappear as one continuously changes the parameters of the compactification. Spinors in higher dimensions decompose as left–right symmetric pairs with respect to four dimensions, but for suitable compactification manifolds, it is possible to obtain chiral fermions. However, it turns out to be impossible

to obtain chiral fermions in the required representations of the Standard Model group. We will see, though, that string theory can generate both gauge groups and chiral fermions upon compactification.

25.2 Closed strings on tori

So far we have considered compactifications of field theories in higher dimensions. But general higher-dimensional field theories are non-renormalizable, and must be viewed as low-energy limits of some other structure. The only sensible structure we know in higher dimensions is string theory. At the same time, if string theory is to have anything to do with the world around us, it must be compactified to four dimensions.

It is not complicated to repeat this analysis for the case of closed strings on circles, or more generally on tori. Consider first compactifying one dimension, X^9, on a circle of radius $2\pi R$. We require that states be invariant under translations by $2\pi R$. This means that the momenta, as in the field theory case, are quantized,

$$p^9 = \frac{n}{R}. \tag{25.14}$$

But now there is a new feature. Because of the identification of points, the string fields themselves (X^9) need not be strictly periodic. Instead, we now have the mode expansion:

$$X^9 = x^9 + p^9\tau + 2mR\sigma + \frac{i}{2}\sum_{n\neq 0}\frac{1}{n}\left(\alpha_n^9 e^{-in(\tau-\sigma)} + \tilde{\alpha}_n^9 e^{-in(\tau+\sigma)}\right), \tag{25.15}$$

where m is an integer. The states with non-zero m are called "winding" states. They correspond to the possibility of a string winding around, or wrapping, the extra dimension. Now the mass operator, in addition to including a contribution $(p^9)^2 = n^2/R^2$, includes, as well, a contribution from the windings, m^2R^2 (if there is no momentum). If R is large compared with the string scale, these states are very heavy. At small R, however, these states become light, while the momentum (Kaluza–Klein) states become heavy. This reciprocity often corresponds, as we will see, to a symmetry between compactification at large and at small radius.

Let's focus on the various superstring theories. It is convenient to break up X^9 in terms of left- and right-moving fields:

$$X_{\rm L}^9 = \frac{x^9}{2} + \left(\frac{n}{2R} + mR\right)(\tau-\sigma) + \frac{i}{2}\sum_{n\neq 0}\frac{1}{n}\alpha_n^9 e^{-in(\tau-\sigma)} \tag{25.16}$$

$$X_{\rm R}^9 = \frac{x^9}{2} + \left(\frac{n}{2R} - mR\right)(\tau+\sigma) + \frac{i}{2}\sum_{n\neq 0}\frac{1}{n}\tilde{\alpha}_n^9 e^{-in(\tau+\sigma)}. \tag{25.17}$$

It is then natural to define left- and right-moving momenta:

$$p_{\rm L} = \frac{n}{2R} + mR \quad p_{\rm R} = \frac{n}{2R} - mR. \tag{25.18}$$

The world-sheet fermions are untouched by this compactification. The mass operators are essentially as before, with p replaced by $p_{\rm L}$ for the left movers and $p_{\rm R}$ for the right movers:

$$L_0 = \frac{1}{2}p_{\rm L}^2 + N \quad \tilde{L}_0 = \frac{1}{2}p_{\rm R}^2 + \tilde{N}. \tag{25.19}$$

Suppose we compactify on a simple product of circles. The left- and right-moving momenta form a lattice:

$$p_{\rm L}^I = \frac{n^I}{R^I} + 2m^I R^I \quad p_{\rm R}^I = \frac{n}{R^I} - 2m^I R^I. \tag{25.20}$$

Let's determine the spectrum, focusing on the light states. Consider, first, the heterotic string, and to simplify the formulas, we take the $O(32)$ case. The $O(32)$ symmetry is unbroken. The original ten-dimensional gauge bosons,

$$\left|A_M^{AB}\right\rangle = \lambda^A_{-1/2}\lambda^B_{-1/2}\psi_{M\ -1/2}|p\rangle, \tag{25.21}$$

now decompose into a set of four-dimensional gauge bosons, corresponding (in light cone gauge) to $M = 2, 3$, and six scalars, $M = I$. The graviton, scalar, and antisymmetric tensor field now decompose as a set of scalars, g_{IJ}, B_{IJ}, vectors $g_{\mu i}$, $B_{\mu I}$, a four-dimensional graviton, $g_{\mu\nu}$, antisymmetric tensor, $b_{\mu\nu}$, and scalar, ϕ.

To understand space-time fermions, work in light cone gauge, and return to our description of $O(8)$ spinors. Group the γ-matrices into a set associated with the internal six dimensions, and one associated with the (transverse) Minkowski directions. In other words, instead of the four creation and annihilation operators, a^i, $a^{\bar{i}}$, we group these into one set of three (labelled a^i, where now $i = 1, 2, 3$, and b, and their conjugates). So the $8_{\rm s}$, which previously consisted of the states

$$|0\rangle \quad a^{i\dagger}a^{j\dagger}|0\rangle \quad a^{1\dagger}a^{2\dagger}a^{3\dagger}a^{4\dagger}|0\rangle, \tag{25.22}$$

now decomposes as:

$$|0\rangle \quad a^{i\dagger}a^{j\dagger}|0\rangle \quad b^{\dagger}a^{j\dagger}|0\rangle \quad b^{\dagger}a^{1\dagger}a^{2\dagger}a^{3\dagger}|0\rangle. \tag{25.23}$$

There are four states with no bs, and four with one b. These groups have opposite *four-dimensional* helicity. They can also be classified according to their transformation properties under $O(6)$. $O(6)$ is isomorphic to the group $SU(4)$. We have just seen that $8_{\rm s} = 4 + \bar{4}$. We can also see that under the $SU(3)$ subgroup of $SU(4)$, the spinor decomposes as

$$8 = 3 + \bar{3} + 1 + 1. \tag{25.24}$$

Consider how the gravitino in ten dimensions decomposes under $O(3,1) \times SU(4)$. We see that the gravitino consists of a set of spin-3/2 particles in the 4 of $SU(4)$, and their antiparticles. So, from the perspective of four dimensions, this is a theory with $N = 4$ supersymmetry. This is not really surprising since the ten-dimensional theory was a theory with 16 supercharges, and none of these are touched by this reduction to four dimensions.

Because of the high degree of SUSY, one cannot write a potential for the scalar fields, g_{IJ}, b_{IJ}, etc.; they are exact flat directions. If we redo our Casimir energy calculation, we will find that, because there is a fermionic state degenerate with every bosonic state, there is a cancellation.

What do these moduli correspond to? Those which arise from the diagonal components of the metric correspond to the fact that the radii are not fixed. There is a string solution for any value of the R^I. The off-diagonal components are related to the fact that the general torus in six dimensions is not simply a product of circles; there can be non-trivial angles.

The massless scalars arising from the gauge bosons, A^I, are also moduli. For constant values of these fields, there is no associated field strength, so they carry zero energy. But there are non-trivial Wilson lines:

$$U_I = e^{i \int_0^{2\pi R_I} dx^I A_I}. \tag{25.25}$$

Because of the periodicity, these are gauge-invariant, and correspond to distinct physical states. These moduli are often themselves called Wilson lines.

The periodicities of a general N-dimensional torus can be characterized in terms of N basis vectors, e_a^I, $a = 1, \ldots, N$. The theory is defined by the identifications:

$$X^I = X^I + 2\pi n^a e_a^I. \tag{25.26}$$

The set of integers define a lattice. To determine the allowed momenta, we define the dual lattice, with unit vectors: $\tilde{e}_a^I$, satisfying:

$$\tilde{e}_a^I e_b^I = \delta_{a,b}. \tag{25.27}$$

In terms of these, we can write the momenta for the general torus:

$$p^I = n^a \tilde{e}_a^I, \tag{25.28}$$

while the windings are:

$$w^I = m^a e_a^I. \tag{25.29}$$

We can break these into left-moving and right-moving parts:

$$p_{\rm L}^I = (p^I/2 + w^I) \quad p_{\rm R}^I = (p^I/2 - w^I). \tag{25.30}$$

The lattice of left- and right-moving momenta, $(p_{\rm L}, p_{\rm R})$, has some interesting features. Thought of as a Lorentzian lattice it is even and self-dual. The "even" refers to the fact that the inner product of a vector with itself:

$$p_{\rm L}^2 - p_{\rm R}^2 = 2nm, \tag{25.31}$$

is even. The self-duality means that the basis vectors of the lattice and the dual are the same (Eq. (25.27)).

In bosonic or Type II theories, these are are the most general four-dimensional compactifications with $N = 8$ supersymmetry. The different possible choices of torus define a moduli space of such theories. These moduli correspond to varying the metric and antisymmetric tensor fields. In the heterotic case, the four dimensional theory has $N = 4$ supersymmetry. Additional moduli arise from Wilson lines. As for the simple compactification on a circle, these are essentially constant gauge fields. A constant gauge field is almost a pure gauge transformation (take I fixed, for simplicity):

$$A^I = ie^{ix_I A^I} \partial^I e^{-ix_I A^I} = ig\partial^I g^\dagger \tag{25.32}$$

but the gauge transformation is only periodic if $A^I = 1/R_I$. In this case, the Wilson line is unity. But we *can* do a redefinition of all of the charged fields, which eliminates the A^Is:

$$\phi = g\phi'. \tag{25.33}$$

With this choice, charged fields are no longer periodic, but obey boundary conditions:

$$\phi\prime(X^I) = e^{2\pi i R_I A^I} \phi\prime. \tag{25.34}$$

This means that the momenta are shifted:

$$p^I = \frac{n}{R_I} + A^I. \tag{25.35}$$

Shortly, we will see how all of the different momentum lattices can be understood in terms of constant background fields.

25.3 Enhanced symmetries

For large radius, the spectrum of the toroidally compactified string theory is very similar to that expected from Kaluza–Klein field theories. The principal new feature, the winding states, is not important. At smaller radius, however, these states introduce startling new phenomena. We focus, first, on compactification of just one

dimension. Examining the momenta

$$p_{\rm L} = \frac{m}{2R} + nR \quad p_{\rm R} = \frac{m}{2R} - nR \tag{25.36}$$

we see that these are symmetric under $R \to 1/(2R)$. This symmetry is often called "T-duality." It means that there is not a sense in which one can take the compactification radius to be arbitrarily small; it is our first indication that there is some sort of fundamental length scale in the theory. T-duality is *not* a feature of the compactification of field theory; the string windings are critical.

What is the physical significance of this symmetry? The answer depends on which string theory we study. Consider the heterotic string. We first ask whether duality is truly a symmetry, or just a feature of the spectrum. To settle this, we can check that it has a well-defined action on all vertex operators. Alternatively, we note that there is a self-dual point: $R_{\rm sd} = 1/\sqrt{2}$. Examining Eq. (25.19) we see that, at this radius, various states can become massless. These include both scalars (from the point of view of the non-compact dimensions) and gauge bosons:

$$\psi^{I,\mu}_{-1/2}|n = \pm 1, m = \mp 1\rangle. \tag{25.37}$$

Together with the $U(1)$ gauge boson, the spin-one particles form the adjoint of an $SU(2)$. We can check this by studying the operator product expansions of the associated vertex operators (see the exercises at the end of this chapter).

Now we can understand the $R \to 1/R$ symmetry. At the fixed point, the symmetry is an unbroken symmetry. It transforms:

$$p_{\rm L} \to -p_{\rm L} \quad p_{\rm R} \to p_{\rm R}. \tag{25.38}$$

In world sheet terms, this corresponds to a change of sign of $\partial X_{\rm L}$:

$$\partial X_{\rm L} \to -\partial X_{\rm L} \quad \partial X_{\rm R} \to \partial X_{\rm R}. \tag{25.39}$$

From (25.37), $X_{\rm L}$ is the third component of isospin, T_3 so $T_3 \to -T_3$ under T-duality.

This transformation is a 90° rotation about the 1 or 2 axis in the $SU(2)$ space, i.e. it is a gauge transformation! This means that the large and small radii not merely exhibit the same physics, they *are* the same. It also means that, provided the theory makes sense, the symmetry is an exact symmetry of the theory, in perturbation theory and beyond. As for any gauge symmetry, any violation of the symmetry would signal an inconsistency.

Returning to the self-dual point, the momentum lattice at this point can be thought of as a group lattice, with the $p_{\rm L}$s labeling the $SU(2)$ charges. Much larger symmetry groups can be obtained by making special choices of the torus, Wilson lines and antisymmetric tensor fields.

In other string theories the symmetry has a different significance. Consider the Type II theories; take the case of IIA for definiteness. Then since $\psi_R^9 \to -\psi_R^9$, the GSO projection in the right-moving Ramond sectors is flipped. So this transformation takes the Type IIA theory to the Type IIB theory. In other words, the IIA theory at large R is equivalent to the IIB theory at small R.

25.4 Strings in background fields

The possibilities for string compactification are much richer than tori, and we will explore them in this and the next chapter. We can approach the problem in two ways, each of which is very useful. First, we can examine the low-energy effective field theory which describes the massless modes of the string in ten dimensions, and look for solutions corresponding to large internal spaces. The effective action can be organized into terms with more and more derivatives. The spaces must be large in order that this use of the low-energy effective action makes sense. Alternatively, we can look for more direct ways to construct classical solutions in string theory. Both approaches have turned out to have great value.

We will first formulate the string problem in a more general way. We want to ask: how do we describe a string propagating in a background which is not flat? The background might be described by a metric, G_{MN}, but it might also include an antisymmetric tensor, B_{MN}, a dilaton, ϕ, and, in the case of the heterotic string, gauge fields. Let's first focus on the metric. Start with the bosonic string. It is natural to generalize the string action:

$$\frac{1}{2\pi}\int d^2\sigma\, \partial_\alpha X^M \partial^\alpha X^N \eta_{MN} \tag{25.40}$$

to

$$\frac{1}{2\pi}\int d^2\sigma\, \partial_\alpha X^M \partial^\alpha X^N G(X)_{MN}. \tag{25.41}$$

From a world sheet point of view, we have replaced a simple free field theory with a non-trivial, interacting field theory. We can think of the X^Ms as fields which propagate on a manifold with metric G_{MN}. Often this space is called the "target space" of the theory; the Xs then provide a mapping from two-dimensional spacetime to this target space.

This looks plausible, but we can give some evidence that it is the correct prescription. Suppose, in particular, we consider a metric which is nearly that of flat space:

$$G_{MN} = \eta_{MN} + h_{MN}. \tag{25.42}$$

Substitute this form in the action, and examine the path integral for the field theory:

$$Z[h] = \int [dX^M] e^{iS_0 + \frac{1}{2\pi}\int d^2\sigma \partial_\alpha X^M \partial^\alpha X^N h(X)_{MN}}. \tag{25.43}$$

Differentiating with respect to h brings down a vertex operator for the graviton. In other words, the path integral for this action is the generating functional for the graviton S-matrix.

This observation suggests a general treatment for backgrounds for the massless particles

$$I = \frac{1}{2\pi}\int d\tau \int_0^\pi d\sigma (g_{IJ}\partial_\alpha X^I + \epsilon^{\alpha\beta} B_{IJ}\partial_\alpha X^I \partial_\beta X^J). \tag{25.44}$$

The corresponding path integral generates the S-matrix elements for both the graviton and the antisymmetric tensor field. But we would like to consider configurations which are not close to the flat metric with vanishing B_{MN}. We can ask: what are acceptable backgrounds for string propagation? To answer this question, we need to remember that for the free string, *conformal invariance* was the crucial feature to the consistency of the picture. It was conformal invariance which guaranteed Lorentz invariance and unitarity. So we need to look for interacting two-dimensional field theories which are conformally invariant.

25.4.1 The beta function

Field theories of the type we have just encountered are called non-linear sigma models. In $1+1$ dimensions, these are renormalizable theories: g_{IJ}, B_{IJ}, etc., are dimensionless. A priori, however, they are general functions of the fields, and there are an infinite – continuously infinite – set of possible couplings.

Physically, the statement that these theories must be conformally invariant is the statement that their beta functions must vanish. To get some feeling for what this means, let's consider a special situation. Suppose that B_{IJ} vanishes, and that the metric is close to the flat space metric:

$$g_{MN} = \eta_{MN} + \int d^D k\, h_{MN}(k) e^{ik\cdot x}. \tag{25.45}$$

The action is then:

$$I = \frac{1}{2\pi}\int d^2\sigma \left(\eta_{IJ}\partial_\alpha X^I \partial^\alpha X^J + \sum_k h_{IJ}(k) e^{ik\cdot x} \partial_\alpha X^I \partial^\alpha X^J \right). \tag{25.46}$$

We can treat the term involving h as a perturbation. Working to second order, we have:

$$\left\langle \int d^2z_1 (h_{\mu\nu}(k) e^{ik\cdot X(z_1)} \partial X(z_1)^\mu \partial X(z_1)^\nu) \times \int d^2z_2 \left(h_{\rho\sigma}(k') e^{ik'\cdot X(z_2)} \partial X(z_2)^\rho \partial X(z_2)^\sigma \right) \right\rangle. \tag{25.47}$$

Let's write this simply as:

$$\int d^2z \int d^2z' h_1 \mathcal{O}_1(z_1) h_2 \mathcal{O}_2(z_2). \tag{25.48}$$

Ultraviolet divergences will arise in this integral when $z_1 \to z_2$. In this limit, we can use the operator product expansion,

$$\mathcal{O}_1(z_1)\mathcal{O}_2(z_2) = \frac{c_{12j}}{|z_1 - z_2|^2} \mathcal{O}_j(z_2) + \cdots. \tag{25.49}$$

The integral over z_2 is ultraviolet divergent. If we cut it off at scale Λ^{-1}, we have the correction to the world sheet Lagrangian:

$$\int d^2z \; h_1 h_2 c_{12j} \; \mathcal{O}_j \; \ln(\Lambda). \tag{25.50}$$

There is another divergence associated with the couplings h_1 and h_2; this comes from normal ordering. In the case of the graviton vertex operator, if we simply expand the exponential factors and contract the xs, we obtain:

$$\int d^2z h_1(x) k^2 \ln(\Lambda). \tag{25.51}$$

Requiring, then, that the beta function for the coupling h_1 vanishes, gives:

$$k^2 h_1 + h_2 h_3 \, c_{123} = 0. \tag{25.52}$$

Recall, now, that c_{ijk} is the three-point coupling for the three fields. So this is just the equation of motion, to quadratic order in the fields.

This result is general. At higher orders, one encounters divergences of two types. First, there are terms involving a single logarithm of the cutoff, times more powers of the fields. Second, there are terms involving higher powers of logarithms. The higher powers are, from a renormalization perspective, associated with iterations of lowest order divergences, and they are systematically subtracted off in computing the beta functions. From a space-time point of view, these correspond to the appearance of massless intermediate states, which must be subtracted off in constructing the effective action or equations of motion.

This procedure can be used to recover Einstein's equations. A more elegant and efficient approach is to apply the background field method. For a general gravitational background, one can view X as a fixed background, which solves the

two-dimensional equations of motion, and study fluctuations about it. For a suitable choice of coordinates, the metric is second order in the fluctuations. One can include in this analysis background antisymmetric tensor fields and a background dilaton. The antisymmetric tensor can be analyzed along the lines of our analysis of $h_{\mu\nu}$. The dilaton is more subtle. In our action above, we omitted one possible coupling: the two-dimensional curvature. The dilaton couples to the world sheet fields through:

$$\int d^2\sigma \Phi \mathcal{R}^{(2)}. \tag{25.53}$$

The full analysis leads to the equations of motion:

$$\beta_{\mu\nu} = 0 = \alpha' R_{\mu\nu} + 2\alpha' \nabla_\mu \nabla_\nu \Phi - \frac{\alpha'}{4} H_{\mu\lambda\omega} H_\nu^{\lambda\omega} \tag{25.54}$$

$$\beta^B_{\mu\nu} = -\frac{\alpha'}{2} \nabla^\omega H_{\omega\mu\nu} + \alpha' \nabla^\omega \Phi H_{\omega\mu\nu} + \mathcal{O}(\alpha')^2 \tag{25.55}$$

$$\beta^\Phi = \frac{D-26}{6} - \frac{\alpha'}{2} \nabla^2 \Phi + \alpha' \nabla_\omega \Phi \nabla^\omega \Phi - \frac{\alpha'}{24} H^{\mu\nu\lambda} H_{\mu\nu\lambda}. \tag{25.56}$$

It is possible to extend these methods to describe quantum corrections to the equations, at least in the case of supersymmetric compactifications.

25.4.2 More general tori

As a first application, we consider the heterotic string theory in the case of more general tori.

For general metric and backgrounds for both the antisymmetric tensor and gauge fields, one obtains a somewhat more involved expression for the momenta. A particularly elegant way to derive this is to argue that constant background fields should effect only slow modes of the string. In the presence of background, constant metric and antisymmetric tensor fields, the action is:

$$I = \frac{1}{2\pi} \int d\tau d\sigma \int_0^\pi (g_{IJ} \partial_\alpha X^I \partial^\alpha X^j + \epsilon^{\alpha\beta} B_{IJ} \partial_\alpha X^I \partial_\beta X^J). \tag{25.57}$$

To realize the notion of slowly varying fields, one makes the Ansatz:

$$X^I = q^I(\tau) + 2\sigma m^I \tag{25.58}$$

where the second term allows for the possibility of winding. Substituting this back in the action and performing the integral over σ:

$$I = \int d\tau \left(\frac{1}{2} g_{IJ} \dot{q}^I \dot{q}^J + 2 B_{IJ} \dot{q}^I m^J - 2 g_{IJ} n^I n^J \right). \tag{25.59}$$

Now we can read off the canonical momenta:

$$P_I = g_{IJ}\dot{q}^J + 2B_{IJ}m^J. \tag{25.60}$$

In quantum mechanics, it is the canonical momenta which act by differentiation on wave functions, so it is the canonical momenta which must be quantized for a periodic system:

$$P_I = n_I. \tag{25.61}$$

In terms of q^I, this gives:

$$\dot{q}^I = g^{IJ}m_J - 2B^I_J n^J. \tag{25.62}$$

Finally, integrating this equation and substituting back into X^I:

$$X^I = q^I + 2\sigma m^I + \tau\left(g^{IJ}n_J - 2B^I_J m^J\right). \tag{25.63}$$

From this, we can read off the left- and right-moving momenta:

$$\begin{aligned} p^I_{\mathrm{L}} &= m^I + \frac{1}{2}g^{IJ}n_J - g^{IJ}B_{JK}m^K \\ p^I_{\mathrm{R}} &= -m^I + \frac{1}{2}g^{IJ}n_J - g^{IJ}B_{JK}m^K. \end{aligned} \tag{25.64}$$

Once again, $p_{\mathrm{L}} \cdot p'_{\mathrm{L}} - p_{\mathrm{R}} \cdot p'_{\mathrm{R}}$ is an integer; the lattice, thought of as a Lorentzian lattice, is even and self-dual.

Including Wilson lines is slightly more subtle, because of their asymmetric coupling between left and right movers. For small A, the modification is essentially what we guessed above. There is also a modification of the internal, E_8 charge lattice.

25.5 Bosonic formulation of the heterotic string

We have seen that in toroidal compactifications of string theory, new unbroken gauge symmetries can arise at particular radii. We have also seen that a toroidal compactification can be described by a lattice. So far, in describing the heterotic string, we have worked in what is known as the fermionic formulation. There is an alternative formulation, in which the 32 left-moving fermions are replaced by 16 left-moving bosons.

It is an old result that two-dimensional fermions are equivalent to bosons; more precisely, two real, left-moving fermions are equivalent to a single real boson, and vice versa. The correspondence, for a complex fermion, λ, is:

$$\lambda(z) = e^{i\phi(z)}, \tag{25.65}$$

where ϕ is a left-moving boson. The equal sign here is subtle; at finite volume, care is required with the zero modes, as we will see. To be convinced that this equivalence is plausible, consider correlation functions at infinite volume. From our previous analyses of two-dimensional Green functions, we have:

$$\langle \lambda(z)\lambda(w)\rangle = \left\langle e^{i\phi(z)}e^{i\phi(w)}\right\rangle \sim \frac{1}{z-w}. \tag{25.66}$$

This suggests that, say, in the case of the $SO(32)$ heterotic string, we can replace the 32 left-moving fermions by 16 left-moving bosons. Note that this means, loosely, that we have 26 left-moving coordinates, as in the bosonic string (but still only 10 right-moving bosons). At finite volume (i.e. $0 < \sigma < \pi$), we can write the usual mode expansions for these fields:

$$X_{\rm L}^A = \frac{1}{2}p_{\rm L}^A + \frac{i}{2}\sum_n \frac{1}{n}\tilde{\alpha}_n e^{-in(\tau+\sigma)}. \tag{25.67}$$

Now the $p_{\rm L}$s are elements of the group lattice. Modular invariance requires that the lattice be even and self-dual. In 16 dimensions, there are two such lattices, those of $O(32)$ and $E_8 \times E_8$.

The bosonization of fermions which we have described here is useful for the right-moving fields as well and also for the fermions of the Type II theories. We have avoided discussing space-time supersymmetry in the RNS formalism because the fermion vertex operators and the supersymmetry generators must change the boundary conditions on two dimensional fields. But in this bosonized form, this problem is simpler. Once again, we have relations of the form

$$\psi_i \sim e^{i\phi_i}. \tag{25.68}$$

The ϕs live on a torus, whose "momenta" describe both N and RS states. Operators of the form $e^{\frac{i\phi}{2}}$ change NS to R states, i.e. they connect bosons to fermions. This connection allows construction of fermion vertex operators and the supersymmetry generators.

25.6 Orbifolds

Toroidal compactifications of string theory are simple; they involve free two-dimensional field theories. But they are also unrealistic. Even in the case of the heterotic string, they have far too much supersymmetry and their spectra are not chiral. There is a simple construction which reduces the amount of supersymmetry, yielding models with interesting gauge groups and a chiral structure. The corresponding world sheet theories are still free, so explicit computations are straightforward. These constructions are also interesting in other ways. They correspond

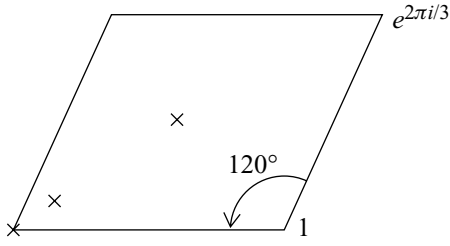

Fig. 25.1. Torus that admits a Z_3 symmetry, allowing an orbifold construction.

to particular submanifolds of the moduli space of larger classes of solutions. They exhibit interesting features like discrete symmetries and subtle cancellations of four-dimensional anomalies. At low orders, it is a simple matter to work out their low-energy effective actions. Through a combination of world sheet and space-time methods, one can understand their perturbative and in some cases non-perturbative dynamics.

In this chapter, we will work out one example in some detail. Other examples can be studied in a similar way. We will also mention some other free field constructions of interesting string solutions.

We start with a toroidal compactification on a particular lattice, a product of three tori as shown in Fig. 25.1. It is convenient to introduce complex coordinates,

$$z^1 = x^1 + ix^2 \quad z^2 = x^3 + ix^4 \quad z^3 = x^5 + ix^6. \tag{25.69}$$

This lattice is invariant under a Z_3 symmetry

$$z^i \to e^{\frac{2\pi i}{3}} z^i. \tag{25.70}$$

This can be seen by carefully examining the figure. The lattice vector $(1, 0)$, for example, in the original cartesian coordinates is rotated into the lattice vector $(-1/2, 1/\sqrt{2})$. This is related by translation by a lattice vector to $(1/2, 1/\sqrt{2})$.

Now we identify points under the symmetry, i.e. two points related by a symmetry transformation are considered to be the same point. The result is almost a manifold, but not quite. There are three points which are invariant under the symmetry. These are called "fixed points." These are the points:

$$(0, 0);\ (1/2, \sqrt{3}/2);\ (1, \sqrt{3}). \tag{25.71}$$

The geometry near each of these points is singular. If one parallel transports about, say the point at the origin, after 120°, one returns where one started. The space is said to have a deficit angle. It is as if there was an infinite amount of curvature located at each of the points. Such a space is called an "orbifold."

In quantum mechanics, requiring such an identification of points under a symmetry means requiring that states be invariant under the quantum mechanical operator which implements the symmetry. Consider the various states of the original ten-dimensional theory. In the Type II theory, for example, in the NS–NS sector, we have the states, before making any identifications:

$$\tilde{\psi}^{\mu}_{-1/2}\psi^{\nu}_{-1/2}|0\rangle \quad \tilde{\psi}^{\bar{j}}_{-1/2}\psi^{i}_{-1/2}|0\rangle \quad \tilde{\psi}^{j}_{-1/2}\psi^{\bar{i}}_{-1/2}|0\rangle \tag{25.72}$$

$$\tilde{\psi}^{j}_{-1/2}\psi^{i}_{-1/2}|0\rangle \quad \tilde{\psi}^{\bar{i}}_{-1/2}\psi^{\bar{j}}_{-1/2}|0\rangle. \tag{25.73}$$

After the identification, the first set of states are invariant; the latter are not. These states all have simple interpretations. The first are the four dimensional graviton, antisymmetric tensor and dilaton. The second are the moduli of the torus. The parts symmetric under $i \to \bar{j}$ correspond to the metric components in the original theory $g_{i\bar{j}}$. The antisymmetric parts correspond to the corresponding components of $B_{i\bar{j}}$.

The diagonal components, $g_{i\bar{i}}$, are easily understood. Changing slightly the value of these components correspond to changing the overall radius of the ith torus. This does not change the symmetry properties. The off-diagonal components, $g_{1\bar{2}}$, etc., correspond to deformations which mix up the three planes, but leave a lattice with an overall Z_3 symmetry.

To understand what happens to the supersymmetries, focus on the gravitino. It is convenient to work in light cone gauge, and to decompose the spinors as we did earlier. To determine how the spinors transform under the Z_3, we need to decide how the state we called $|0\rangle$ transforms under the symmetry. Consider a rotation, say in the 12 plane, by 120°. The rotation generator is:

$$S_{12} = \frac{i}{4}(\gamma_1\gamma_2 - \gamma_2\gamma_1) = a^{1\dagger}a^1 + 1/2. \tag{25.74}$$

So the rotation of the state $|0\rangle$, is described by

$$e^{\frac{2\pi i}{6}s_{12}}|0\rangle = e^{\frac{2\pi i}{6}}|0\rangle. \tag{25.75}$$

The transformations of the other states can then be read off from the transformation laws of the a^is:

$$|0\rangle \to e^{\frac{-2\pi i}{6}}|0\rangle \quad a^{\bar{i}}|0\rangle \to e^{\frac{2\pi i}{6}}a^{\bar{i}}|0\rangle. \tag{25.76}$$

Now we have to be a bit more precise about the orbifold action. This is a product of Z_3s for each of the planes. But we see that acting on fermions, the separate transformations are Z_6s. In order that the group action be a sensible Z_3, we need to take, for example:

$$Z^1 \to e^{\frac{2\pi i}{3}}Z^1 \quad Z^2 \to e^{\frac{2\pi i}{3}}Z^2 \quad Z^3 \to e^{\frac{-4\pi i}{3}}Z^3. \tag{25.77}$$

With this definition, the fermion component $0 \leftrightarrow |0\rangle$ is invariant under the orbifold projection. The components $i \leftrightarrow a^{\bar{i}}|0\rangle$ are not.

We can label the gravitinos:

$$\psi^{\mu}_{0,\alpha} \quad \tilde{\psi}^{\mu}_{\tilde{0},\alpha} \quad \psi^{\mu}_{i,\alpha} \quad \tilde{\psi}^{\mu}_{i,\alpha}. \tag{25.78}$$

After the projection, instead of eight gravitinos, as in the toroidal case, there are only two; we have $N = 2$ supersymmetry in four dimensions.

In addition to projecting out states, we need to consider a new class of states. We can consider closed strings which sit at the fixed points. More precisely, in addition to the strict periodic boundary condition, we can consider strings which satisfy:

$$X^i(\sigma + \pi) = e^{\frac{2\pi i}{3}} X^i(\sigma). \tag{25.79}$$

These boundary conditions do not permit the usual bosonic zero modes. Instead, we have a mode expansion:

$$X^i = x^i_{(a)} + \frac{i}{2} \sum_n \left(\alpha^i_{n-\frac{1}{3}} e^{2i(n-\frac{1}{3})(\sigma-\tau)} + \tilde{\alpha}^i_{n-\frac{1}{3}} e^{2i(n-\frac{1}{3})(\sigma+\tau)} \right). \tag{25.80}$$

The mode numbers are now fractional; the absence of a momentum term indicates that the strings sit at the fixed points (labeled by a). In this case, there are 27 fixed points. For the fermions, we again have to distinguish Ramond and Neveu–Schwarz sectors. In the NS sectors, the fermions have modes which differ from integers by multiples of $1/2 - 1/3 = 1/6$:

$$\psi^i = \sum \psi_{n-\frac{1}{6}} e^{-2i(n-\frac{1}{6})(\tau-\sigma)} \tag{25.81}$$

with a similar expansion for $\tilde{\psi}$.

We can readily work out the normal ordering constant, using the formula we wrote earlier (Eq. (22.30)). We have, in the NS–NS sector:

$$a = 6 \times \frac{1}{4}(1/3 \times 2/3) - 6 \times \frac{1}{4}(1/6) \times (5/6) - 4 \times \frac{1}{4}(1/2) \times (1/2) = 0. \tag{25.82}$$

So the ground state is massless in the twisted sectors. Again, because of $N = 2$ supersymmetry, there can be no potential for this field. So there is a modulus in each twisted sector. Unlike the moduli in the untwisted sector, this modulus does not correspond to simply changing the features of the torus which defines the orbifold. Instead, it represents a deformation which, from a space-time viewpoint, smooths out the orbifold singularity. The resulting smooth space is an example of a Calabi–Yau manifold, of the type we will discuss in the next chapter.

Let's turn to the heterotic string theory on this orbifold. We will take the same projector on the spatial coordinates, X^i, as before. As a result, there is only one

gravitino; the four-dimensional theory has $N = 1$ supersymmetry. The moduli are in one to one correspondence with the scalars of the NS–NS sector of the $N = 2$ theory: $g_{i\bar{j}}, B_{i\bar{j}}, \phi$. We can also make a projection on the world sheet gauge degrees of freedom. There are many possible choices of this gauge transformation; the principal restriction comes from the requirement of modular invariance. A particularly simple one is almost symmetrical between the left and right movers. In the fermionic formulation it works as follows. Take $E_8 \times E_8$ for definiteness. Of the 16 fermions in the first E_8 single out 6, and rewrite them in terms of three complex fermions, λ^i. Call the remaining ten fermions λ^a. Now, in the projection, require invariance under

$$Z^i \to e^{\frac{2\pi i}{3}} Z^i \quad \psi^i \to e^{\frac{2\pi i}{3}} \psi^i \quad \lambda^i \to e^{\frac{2\pi i}{3}} \lambda^i. \tag{25.83}$$

In the untwisted sector, this projection has no effect on the graviton or the moduli which we have identified previously. But consider the various gauge fields. In ten dimensions, these were vectors in the adjoint of the two E_8s and their fermionic partners. The fields with space-time indices in the internal dimensions now appear as four-dimensional scalars. In order that they be invariant under the full projection, it is necessary to choose their gauge quantum numbers appropriately. In the NS sector for each of the E_8s, the invariant states include the following.

(1) A set of fields in the adjoint of E_6, E_8 and an $SU(3)$. Of these, an $O(10)$ subgroup of the E_6 is manifest in the NS–NS–NS sector, as well as an $O(16)$ subgroup of E_8. Correspondingly, the gauge bosons are:

$$\lambda^a_{-1/2}\lambda^b_{-1/2}\psi^\mu_{-1/2}|0\rangle \tag{25.84}$$

in $O(10)$,

$$\lambda^A_{-1/2}\lambda^B_{-1/2}\psi^\mu_{-1/2}|0\rangle \tag{25.85}$$

in $O(16)$, and, in $SU(3) \times U(1)$:

$$\lambda^i_{-1/2}\lambda^{\bar{i}}_{-1/2}\psi^\mu_{-1/2}|0\rangle. \tag{25.86}$$

Note that all of these states are invariant. The $U(1)$ is actually one of the E_6 generators. E_6 has an $O(10) \times U(1)$ subgroup under which the adjoint representation, which is 78-dimensional, decomposes as:

$$78 = 45_0 + 1_0 + 16_{-1/2} + \overline{16}_{1/2}. \tag{25.87}$$

The remaining E_6 gauge bosons are found in the R–NS–NS sector. The left-moving normal ordering constant vanishes. The ground states in this sector are spinors of $O(10)$, the 16 and $\bar{1}6$ above. The 248-dimensional representation of the second E_8 is filled out as in the uncompactified theory.

(2) Matter fields. These lie in the fundamental representation of E_6, the 27 under $O(10)$. The 27 decomposes as:

$$27 = 1_{-2} + 10_1 + 16_{-1/2}. \tag{25.88}$$

There are nine 10s in the untwisted sectors, corresponding to the states:

$$\lambda^a_{-1/2}\lambda^i_{-1/2}\psi^{\bar{j}}_{-1/2}|0\rangle \tag{25.89}$$

Each of these is one real scalar; we can use the conjugate fields to form nine more real scalars, or eight complex scalars. There are nine singlets of charge -2:

$$\lambda^{\bar{i}}_{-1/2}\lambda^{\bar{j}}_{-1/2}\psi^{\bar{k}}_{-1/2}|0\rangle. \tag{25.90}$$

The 16s come from the R–NS–NS sector.

So we have nine 27s from the twisted sectors, and no $\overline{27}$s; the theory is chiral.

Let's turn now to the twisted sectors. In the Type II case we found moduli in each sector. Here we will find moduli, additional 27s, and more. We first need to compute the normal ordering constants. For the right movers, the calculation is exactly as in the Type II theory, and gives zero. For the left movers in the NS–NS sector, we have:

$$\begin{aligned} a &= -\frac{8}{24} + (6/4)\times(1/3)\times(2/3) - \frac{16}{4}\times\left(-\frac{1}{24}+\frac{1}{4}\right) \\ &\quad + \frac{16}{24} - \frac{10}{4}(1/4) - \frac{6}{4}\times(1/6)\times(5/6) \\ &= -1/2, \end{aligned} \tag{25.91}$$

where the first two terms comes from the bosons, the next two from the fermions in the unbroken E_8, and the last terms from the fermions in the broken E_8. So we can make massless states in a variety of ways.

(1) We can have 10s of $O(10)$:

$$\lambda^a_{-1/2}|0\rangle_{\text{twist}}. \tag{25.92}$$

Note that E_6 invariance requires that this state have $U(1)$ charge $+1$.

(2) Singlet of $O(10)$ with $U(1)$ charge -2:

$$\lambda^{\bar{1}}_{-1/6}\lambda^{\bar{2}}_{-1/6}\lambda^{\bar{3}}_{-1/6}|0\rangle_{\text{twist}}. \tag{25.93}$$

Together with a set of spinorial states from the R–NS sector, this completes a 27 of E_6.

(3) Moduli, other gauge singlets:

$$\alpha^i_{-1/3}\lambda^{\bar{j}}_{-1/6}|0\rangle. \tag{25.94}$$

If we contract the i and $\bar{j}$ index, we find the analog of the twisted sector modulus we had in the Type II theory. The other states represent additional singlets.

All together, then, we have found $9 + 27 = 36$ copies of the 27 of E_6, and 36 moduli. Each 27 comfortably accommodates a generation of the standard model, plus an additional vector-like set of fields. So, while this example is hardly realistic, it is interesting: it predicts a particular number of Standard Model generations, plus additional fields. Whether variants of these ideas can lead to something more realistic is an important question, which we will postpone for the time being.

25.6.1 Discrete symmetries

One of the unappealing features of supersymmetric models as theories of nature is the need to postulate discrete symmetries in order to have a sensible phenomenology. This seems rather ad hoc. One of the features of the orbifold construction we just described is that a variety of discrete symmetries appear naturally. This phenomenon is common, as we will see, in string constructions. Here it is particularly easy to exhibit the symmetries.

We have, for simplicity, considered a particular form for the torus – a particular point in the moduli space, at which the six-dimensional torus is a product of three two-dimensional tori. But at this point (really a surface), there is a large symmetry. First, there is a separate Z_3 symmetry for each plane. (You can check that each plane in fact admits a Z_6 symmetry.) Because of the orbifold projection, one of these acts trivially on all states, but two are non-trivial. If we take the size of each of the three two-dimensional tori to be the same, we also have a permutation symmetry, S_3, among the tori.

The Z_3s are R symmetries. We have already seen that the spinor with index 0 rotates by a phase, $e^{\frac{2\pi i}{6}}$, under the symmetry. By definition, this is an R transformation. This has significant consequences for the low-energy theory, greatly restricting the form of the superpotential.

As an example of the far-reaching consequences of such symmetries, one can show that there are exact flat directions involving the matter fields. Consider the untwisted moduli. One can give expectation values to the $O(10)$ 10 and 1 in one multiplet in a way which respects supersymmetry. Specifically, consider the field, ϕ, corresponding to

$$\phi = \lambda^a_{-1/2}\psi^{\bar{1}}_{-1/2}|0\rangle \tag{25.95}$$

and the corresponding singlet. Both of these are neutral under the rotation in the second plane. So one can not construct any superpotential term involving ϕ alone. One can give an expectation value to the singlet and to the 10 so as to cancel the D terms for E_6. The main danger, then, is a superpotential term of the form:

$$W = \Psi\phi^2 \tag{25.96}$$

with Ψ some other 27. This is E_6 invariant (in terms of $O(10)$ representations, it involves a product of a singlet and two 10s). But no such term is allowed by the discrete symmetries.

So this simple argument shows that the moduli space is even larger than we might have thought. Such symmetries, as they forbid not only certain dimension-four but also dimension-five operators, might also be important to understanding the problem of proton stability and other important phenomenological issues.

The model possesses other symmetries as well. There is Z_3 symmetry under which the twisted sector states transform but the untwisted sector states do not. We will not derive this here, but it is plausible, and can be shown readily if one constructs the vertex operators for the twisted states. Many of the discrete symmetries of the model are subgroups of the Lorentz symmetry of the original higher-dimensional theory. As such, they can probably be thought of as gauge symmetries. This is less obvious for other symmetries, but it is generally believed that the discrete symmetries of string theory all have this character. Searches for anomalies in discrete symmetries, for example, have yielded no examples.

One could ask: why would nature choose a point in the moduli space of some string theory at which there is an unbroken discrete symmetry? At the moment, our understanding of how to connect string theory to nature is not good enough to give a definite answer to this question, but, at the very least, such points are *necessarily* stationary points of the effective potential for the moduli; at the symmetric point, the symmetry forbids linear terms in the action for the charged moduli.

25.6.2 Modular invariance, interactions in orbifold constructions

As in our original string theory constructions, there seems much which is arbitrary in the choices we made above. We also did not spell out what are the appropriate GSO projectors. As for the simple ten-dimensional constructions, the possible GSO projections are constrained by modular invariance. We will leave for the exercises checking some particular cases, but the basic result is easy to state. One can project by any transformation, provided that it has sensible action on fermions and on spinor representations of the gauge group, and provided that one has "level matching" in all of the twisted sectors. This means that one must be able to construct an infinite tower of states in each sector. To understand the significance of this statement, consider a different choice of group action than that we considered above. Instead of twisting by $(1/3, 1/3, -2/3)$, project by $(1/3, -1/3, 0)$. In this case, for example, in the NS–NS–NS sector, the left-moving normal ordering constant is $-13/18$. As a result, one cannot construct any states in the twisted sector which satisfy the level matching condition.

There are other constructions of compactifications with $N = 1$ supersymmetry based on free fields. These include models based purely on free fermions. These models are believed equivalent to orbifold models in which one mods asymmetrically on the left- and right-moving fields. The latter, "asymmetric orbifold," models are interesting in that they potentially have very few moduli. In order to have sensible, unbroken discrete symmetries acting on the left and right, the original lattice typically must sit at a self-dual point. So many moduli are fixed – they are projected out by the orbifold transformation. It is not difficult, in this way, to construct models where there are no moduli neutral under space-time symmetries, except for the dilaton.

25.7 Effective actions in four dimensions for orbifold models

While string theory provides a very explicit set of computational rules, at least for low orders of perturbation theory, these rules are complicated and rather cumbersome. Moreover, except in some special circumstances, we lack a non-perturbative formulation of the theory. Effective field theory methods have proven extremely useful in understanding the dynamics of string theory, both perturbative and non-perturbative. In this section, we will work out the effective action for the orbifold models introduced above. More precisely, we work out the Lagrangian for a subset of the fields, up to and including terms with two derivatives. Many of the features of these Lagrangians will be relevant to the more intricate Calabi–Yau compactifications we will encounter shortly.

In principle, to calculate the effective action, we should calculate the string S-matrix, and write an action for the massless fields which yields the same scattering amplitudes. Alternatively, we can calculate the equations of motion from the beta function and look for an action which reproduces these. But for low order terms in the derivative (α') expansion, for the fields in the untwisted sector, there is a simpler procedure. We know the form of the ten-dimensional effective action; we can simply truncate the theory to four dimensions. To do this, we start by setting all of the charged fields to zero (this includes the gauge fields). We also work at a point with a large discrete symmetry: $Z_3^3/Z_3 \times S_3$. We set all of the fields which transform under these symmetries to zero. This includes all of the moduli, except the overall size of the torus and its superpartners. We then write the metric as:

$$g_{i\bar{j}}(x^\mu) = g_{\bar{j}i}(x^\mu) = e^{\sigma(x^\mu)}\delta_{i\bar{j}}. \tag{25.97}$$

With this parameterization, we are describing the size of the space with respect to a reference metric. We make a similar ansatz for the antisymmetric tensor:

$$b_{i\bar{j}}(x^\mu) = -b_{\bar{j}i}(x^\mu) = b(\sigma(x^\mu))\delta_{i\bar{j}}. \tag{25.98}$$

We must keep also the four-dimensional metric components, $g_{\mu\nu}$ the scalar field ϕ, and the antisymmetric tensor, $B_{\mu\nu}$. Take all of them to be functions of x^μ, the uncompactified coordinates, only. Substituting these fields in the ten-dimensional Lagrangian, Eq. (24.8), the integral over the six internal coordinates is easy, since all fields are independent of the coordinates. One simply obtains $e^{3\sigma(x)}$ from the $\sqrt{g}$ factor. This is just the volume of the internal space, if σ is constant. There are additional factors of $e^{-\sigma}$ coming from the factors of the inverse metric: one from the four-dimensional pieces of the Ricci curvature; one from the kinetic term for ϕ, and three from the $H_{\mu\nu\rho}$ terms. The ten-dimensional curvature term also gives derivative terms in σ. After a short computation, we obtain

$$\mathcal{L} = -\frac{1}{2}e^{3\sigma}R^{(4)} - 3e^{3\sigma}\partial_\mu\sigma\partial^\mu\sigma - \frac{9}{16}e^{3\sigma}\partial_\mu\phi\partial^\mu\phi/\phi^2 - \frac{9}{2}e^{\sigma}\phi^{-3/2}\partial_\mu b\partial^\mu b - \frac{3}{4}\phi^{-3/2}H^{\mu\nu\rho}H_{\mu\nu\rho}. \tag{25.99}$$

It is customary to rescale the metric so that the Einstein term has the standard form:

$$g_{\mu\nu} = e^{-3\sigma}g'_{\mu\nu}. \tag{25.100}$$

After this Weyl rescaling, the action becomes:

$$\mathcal{L} = -\frac{1}{2}R^{(4)} - 3\partial_\mu\sigma\partial^\mu\sigma - \frac{9}{16}(\partial^\mu\phi\partial_\mu\phi)/\phi^2 - \frac{3}{2}e^{-2\sigma}\phi^{-3/2}\partial_\mu b^2 - \frac{3}{4}\phi^{-3/2}e^{6\sigma}H^2_{\mu\nu\rho}. \tag{25.101}$$

It should be possible to cast this Lagrangian as a standard four-dimensional, $N = 1$ supergravity Lagrangian, with a particular Kahler potential. Having thrown away all but a few moduli, there is no superpotential. To determine the Kahler potential, we first note that, in four dimensions, an antisymmetric tensor field is equivalent to a scalar. This follows from counting degrees of freedom; with our usual rules, an antisymmetric tensor in four dimensions has only one degree of freedom. To make this explicit, one performs a "duality transformation" (the word is starting to seem a bit overused)

$$\phi^{-3/2}e^{6\sigma}H_{\mu\nu\rho}(x) = \epsilon_{\mu\nu\rho\sigma}\partial^\sigma a(x). \tag{25.102}$$

The field a is often called the "model-independent axion," because it couples like an axion and its features do not depend on the details of the compactification. Then we define two chiral superfields, whose scalar components are:

$$S = e^{3\sigma}\phi^{-3/4} + 3i\sqrt{2}a \tag{25.103}$$

and

$$T = e^{\sigma}\phi^{3/4} - i\sqrt{2}b. \tag{25.104}$$

Choosing the Kahler potential:

$$K = -\ln(S + S^*) - 3\ln(T + T^*) \tag{25.105}$$

reproduces all of the terms in Eq. (25.101). The reader may want to check all of the terms in this equation carefully, but at the least it is good to make sure one understands how the σ and ϕ dependences are reproduced.

Let's now return to the ten-dimensional gauge field terms, Eq. (24.9). This will allow us to include the matter fields as well as the gauge fields. Rather than consider the full set of fields, we can restrict ourselves to the set which is invariant under each of the separate Z_3s, combined with three separate Z_3s in the gauge group ($\lambda^i \to e^{\frac{2\pi k_i i}{3}}\lambda^i$). This leaves us with three complex scalars, C^i, corresponding to the states

$$C^i \leftrightarrow \lambda^i_{-1/2}\lambda^a_{-1/2}\psi^{\bar{i}}_{-1/2}|0\rangle \tag{25.106}$$

(i is not summed). From the point of view of ten dimensions, these are $A^{ia}_{\bar{i}}$. We also need to include the four-dimensional gauge fields, A^{ab}_{μ}. In this way we obtain the additional terms, after the Weyl rescaling:

$$\mathcal{L}_{\text{gauge}} = -\frac{1}{4}\phi^{-3/4}e^{3\sigma}F^2_{\mu\nu} - 3e^{-\sigma}\phi^{-3/4}D_\mu C^{*\bar{i}}D^\mu C^i + \cdots. \tag{25.107}$$

This can still be put in the standard supergravity form. First, we need to remember that in the duality transformation, $H_{\mu\nu\rho}$ now includes the Chern–Simons terms. Then it is necessary to modify the definition of T to include a contribution from the C fields:

$$T = e^{\sigma}\phi^{3/4} - i\sqrt{2}b + C^{*\bar{i}}C^i \tag{25.108}$$

and to modify the Kahler potential:

$$K = -\ln(S + S^*) - 3\ln(T + T^* - C^*C). \tag{25.109}$$

There is also a coupling of the field S to the gauge fields:

$$\mathcal{L}_S = -\frac{1}{4}SW^2_\alpha. \tag{25.110}$$

This includes a coupling of ϕ and σ to $F^2_{\mu\nu}$, already apparent in Eq. (24.9). The $aF\tilde{F}$ coupling arises from the Chern–Simons term in Eq. (24.10). Recall that

$$H_{\mu\nu\rho} = \partial_{[\mu}B_{\nu\rho]} - \omega_{\mu\nu\rho}. \tag{25.111}$$

So $\int d^4x H^2$, using the definition of a and integrating by parts, gives an $aF\tilde{F}$ coupling. Finally, there is a superpotential cubic in the C fields.

25.7.1 Couplings and scales

It is worth pausing to note the connections between the couplings and scales in different dimensions. We'll focus first on the heterotic string. We see from Eq. (25.110) that S determines the gauge coupling, $S = 1/g^2$. This is as we would naively expect. The ten-dimensional gauge coupling is essentially $1/g_s^2$; when we reduce to four dimensions, the four-dimensional gauge fields correspond to modes which are constant on the internal manifold, so

$$\frac{1}{g_4^2} = \frac{1}{g_s^2} V M_s^6. \tag{25.112}$$

In terms of the fields we defined above, $V = e^{3\sigma}$.

These simple formulas pose a serious problem for the application of weakly coupled heterotic string phenomenology. If we simply identify S with the four-dimensional coupling, then the string coupling satisfies:

$$g_s^2 = g_4^2 V M_s^6. \tag{25.113}$$

So we see that large volume, the limit in which an α' expansion is valid, conflicts with small g_s if g_4 is fixed. We can also write a relation between the string scale and the Planck scale in four dimensions:

$$M_{\rm p}^2 = M_s^8 V g_s^{-2}. \tag{25.114}$$

Solving for M_s and substituting in the previous expression, gives an expression for g_s which is incompatible with weak coupling, if we assume that $V = M_{\rm GUT}^{-6}$.

Later, we will sharpen this strong coupling problem, and consider possible solutions.

25.8 Non-supersymmetric compactifications

So far, we have considered supersymmetric compactifications. This is not necessary, but we will see that non-supersymmetric compactifications raise new conceptual and technical problems.

Perhaps the simplest non-supersymmetric compactification is *Scherk–Schwarz compactification*. Here one compactifies the theory (this can be Type I, Type II, or heterotic) on a torus. In one of the directions, say the ninth direction, one imposes the requirement that bosons should obey periodic boundary conditions, and fermions anti-periodic ones. One can describe this by taking the radius of the extra dimension to be $2 \times 2\pi R$, and performing a projection by

$$P = (-1)^F e^{i(2\pi i)Rp_9}. \tag{25.115}$$

This projection eliminates, for example, the massless gravitinos; there is no supersymmetry, and there is no Bose–Fermi degeneracy in the spectrum. Indeed, in the simplest version, there are no massless fermions at all.

As a result, the usual Fermi–Bose cancellation of supersymmetry does not take place, and at one loop, there is a non-zero vacuum energy. More precisely, there is a potential for the classical modulus R. The calculation of this potential is just the Casimir calculation we encountered earlier. Only the massless ten-dimensional fields contribute; the massive string states give effects which are exponentially suppressed for large R. To see this, one can return to our earlier calculation with a massive state (one of the oscillator excitations of the string). Replacing the sum over integers by an integral in the complex plane and deforming the contour, as before, yields a term exponentially small in the mass. The detailed results depend on the particular model, but typically the potential is negative and goes to zero at large R. In other words, at one loop, the dynamics tends to drive the system to small R. It is not well understood how to study the system beyond one loop.

One can obtain non-supersymmetric theories in four dimensions in many other ways. For example, one can compactify the non-supersymmetric $O(16) \times O(16)$ heterotic theory on a torus (it can also, in some cases, be compactified on the Calabi–Yau spaces discussed in the next chapter). Again, one obtains a one loop potential for the moduli, and the theories are difficult to interpret at the quantum level. Compactifications of the Scherk–Schwarz theories and the non-supersymmetric theories are often related by T-dualities.

Suggested reading

An introduction to Kaluza–Klein theory prior to the development of string theory is provided in the text *Modern Kaluza–Klein Theories* by Appelquist *et al*. (1985). More thorough discussions of aspects of string compactification are provided by the texts of Green *et al*. (1987) and Polchinski (1998). Many of the original papers, particularly the orbifold papers, are highly readable; see, for example, Dixon *et al*. (1986). There are many topics here we have only touched on in this chapter. We gave an argument that vanishing of the beta function of the two-dimensional σ model is equivalent to the equations of motion in space-time, but readers may wish to work through the background field analysis which leads to Einstein's equations. This is described in Polchinski's book and elsewhere. The bosonic formulation of the heterotic string is also well described there, but the original papers are quite readable (Gross *et al*. 1985, 1986). Bosonization and space-time supersymmetry in the RNS formulation are thoroughly discussed by Polchinski; a clear, and rather brief, introduction, is provided by Peskin's 1996 TASI lectures (Peskin, 1997). The non-supersymmetric compactification described here was introduced by Rohm (1984).

Exercises

(1) Derive the gauge terms in the Lagrangian of Eq. (25.7). You can do this by taking the metric to be flat.
(2) Derive the scalar kinetic terms of Eq. (25.8). You can do this by taking the four-dimensional metric, at first, to be flat, and allowing only σ to be a function of x.
(3) Verify, by studying the OPEs of the vertex operators for the different massless fields, that the enhanced symmetry of the bosonic string at the point $R = 1/\sqrt{2}$ is $SU(2) \times SU(2)$. Explain why, in the heterotic string, the symmetry is only $SU(2)$. What is the symmetry in the IIA theory?
(4) For the orbifold model, work out the spectrum in the untwisted sectors in greater detail, paying particular attention to spinorial representations of the O groups, and to the space-time spinors. In particular, make sure you are clear that the 27s are chiral, i.e. all the states in 27s have one four-dimensional chirality, all of those in $\overline{27}$ have the opposite chirality.
(5) Derive the terms in Eq. (25.99) involving $\partial\sigma^2$.
(6) Verify that the Kahler potential of Eq. (25.109) properly reproduces the kinetic terms of the matter fields.

26

Compactification of string theory II. Calabi–Yau compactifications

So far, we have focussed on rather simple models, involving toroidal compactifications and their orbifold generalizations. But while by far the simplest, these turn out to be only a tiny subset of the possible manifolds on which to compactify string theories. A particularly interesting and rich set of geometries is provided by the Calabi–Yau manifolds. These are manifolds which are Ricci flat, $R_{MN} = 0$. Their interest arises in large part because these compactifications can preserve some subset of the full ten-dimensional supersymmetry. This is significant if one believes that low-energy supersymmetry has something to do with nature. It is also important at a purely theoretical level, since, as usual, supersymmetry provides a great deal of control over any analysis; at the same time, there is less supersymmetry than in the toroidal case, so a richer set of phenomena are possible.

This chapter is intended to provide an introduction to this subject. In the first section, we will provide some mathematical preliminaries. Unlike the toroidal or orbifold compactifications, it is not possible, in most instances, to provide explicit formulas for the underlying metric on the manifold and other quantities of interest. The six-dimensional Calabi–Yau spaces, for example, have no continuous isometries (symmetries), so at best one can construct the metrics by numerical methods. But it turns out to be possible to extract much important information without detailed knowledge of the metric from topological considerations. The machinery required to define these spaces and to extract at least some of this information includes algebraic geometry and cohomology theory, subjects not part of the training of most physicists. The following mathematical interlude provides a brief introduction the the necessary mathematics. There is much more in the suggested reading.

26.1 Mathematical preliminaries

Two notions are very useful for understanding Calabi–Yau spaces: differential forms and vector bundles. Differential forms have already appeared implicitly in our

discussion of IIA and IIB string theory. Start with an antisymmetric tensor field, $A_{i_1 i_2 \cdots i_n}$. Suppose that there is a gauge invariance:

$$A_{i_1 \ldots i_n} \to A_{i_1 \ldots i_n} + \frac{1}{n}\left\{\partial_{i_1}\Lambda_{i_2 \ldots i_n} - \partial_{i_2}\Lambda_{i_1 i_3 \ldots i_n} + \cdots (-1)^r \partial_{i_r}\Lambda_{i_1 \ldots i_{r-1} i_{r+1} \ldots i_n}\right\}, \tag{26.1}$$

where Λ is antisymmetric in all of its indices. We can write a shorthand for this:

$$\delta A = d\Lambda, \tag{26.2}$$

where $d\Lambda$ is the "exterior derivative." Acting on an antisymmetric tensor of rank p, the exterior derivative produces a rank $p+1$ antisymmetric tensor, dH:

$$dH_{i_1 \ldots i_{p+1}} = \left(\frac{1}{p+1}\right)\left[\partial_{i_1} H_{i_2 \ldots i_{p+1}} - \partial_{i_2} H_{i_1 i_3 \ldots i_p + 1} + \cdots\right]. \tag{26.3}$$

We can think of this object more abstractly as follows. Antisymmetric tensors with p indices we will call p-forms. A "basis" for the p-forms is provided by the antisymmetrized products of differentials:

$$dx^{i_1} \wedge dx^{i_2} \wedge \cdots \wedge dx^{i_p}. \tag{26.4}$$

We can then write:

$$H = \frac{1}{p!} H_{i_1 \ldots i_p} dx^{i_1} \wedge \cdots \wedge dx^{i_p}. \tag{26.5}$$

The product of two forms, A, B is known as the wedge product, $A \wedge B$. If A is an n-form and B an m-form,

$$(A \wedge B)_{i_1 \ldots i_{n+m}} = \frac{n!m!}{(n+m)!} A_{i_1 \ldots i_n} B_{i_{n+1} \ldots i_{n+m}} + (-1)^P \text{permutations} \tag{26.6}$$

or, more compactly:

$$A \wedge B = \frac{1}{(n+m)!} A_{i_1 \ldots i_n} B_{i_{n+1} \ldots i_{n+m}} dx^1 \wedge \cdots \wedge dx^{n+m}. \tag{26.7}$$

In this language, the exterior derivative can be written as $d \wedge H$ or simply dH, where d is thought of as a one form with components $d_i = \partial_i$.

It is important to practise with this notation, and some exercises are provided at the end of the chapter. One should check that

$$d^2 H = 0. \tag{26.8}$$

It is instructive to write electrodynamics in the language of forms. One should verify that the field strength tensor is a two form, which can be written as

$$F = dA. \tag{26.9}$$

The homogeneous Maxwell's equations (the Bianchi identities for the field strength) follow from $d^2 = 0$,

$$dF = 0. \tag{26.10}$$

Apart from multiplication and differentiation, there is another important operation denoted by $*$, and called the Hodge star. In d dimensions, this takes a p-form to a $d - p$-form:

$$(*H)_{i_1 \ldots i_{d-p}} = \frac{1}{p!} \epsilon^{i_{d-p+1} \ldots i_d}_{i_1 \ldots i_{d-p}} H_{i_{d-p+1} \ldots i_d}. \tag{26.11}$$

A particularly interesting object is $*d$. For example, $*d \wedge d$ is a d-form. But the components of a d-form are necessarily proportional to $\epsilon_{i_1 \ldots i_d}$. With a little work, one can show that:

$$*(*d \wedge d) = \partial^2. \tag{26.12}$$

Using the $*$ operation, we can write the action for a p-form field:

$$S = \frac{1}{2(p+1)!} \int *F \wedge F \tag{26.13}$$

with $F = dA$. This is clearly gauge-invariant. It is easy to check that this reproduces the standard action for electrodynamics.

For physics, we are particularly interested in zero modes of A, i.e. field configurations that satisfy $dA = 0$, but which are not simply gauge transformations, i.e. which cannot everywhere be written as

$$A = d\Lambda. \tag{26.14}$$

A simple example of what is at issue is provided by a gauge field on a circle, $0 \leq y \leq 2\pi R$. The one-form gauge field,

$$A_y = \partial_y \Lambda \quad \Lambda = c\, y \tag{26.15}$$

is not a pure gauge transformation unless $c = n/R$. In electrodynamics, for example, this corresponds to the fact that the Wilson line,

$$U = e^{i \int_0^{2\pi R} dy A_y} \tag{26.16}$$

is gauge-invariant, and non-trivial, again, unless $c = n/R$.

This suggests that we want to consider closed p-forms, α, which satisfy

$$d\alpha = 0 \tag{26.17}$$

but that we are not interested in *exact* forms,

$$\alpha = d\beta. \tag{26.18}$$

More generally, we want to define an equivalence class, known as the cohomology class of α. We will view α and α' as equivalent if

$$\alpha' = \alpha + d\beta, \tag{26.19}$$

where β is well defined everywhere on the manifold.

In general, for field configurations on a manifold M, the number of linearly independent zero modes is known as the Betti number, b_p. This number is related to the number of (basis) p-dimensional submanifolds which are not boundaries of $p+1$-dimensional surfaces. We won't prove this, but we will at least make it plausible. Consider integration of a p-form, α, over a p-dimensional submanifold, Σ:

$$\int_\Sigma \alpha_{i_1 \ldots i_p}\, d\Sigma^{i_1 i_p}. \tag{26.20}$$

By Stokes' theorem, the integral of the exterior derivative of a $p-1$-form, β, over Σ, is related to the integral of β over the boundary of Σ:

$$\int_\Sigma d\beta = \int_{\partial\Sigma} \beta. \tag{26.21}$$

If Σ is compact, it has no boundary, so the integral of $d\beta = 0$.

Two p-forms are in the same cohomology class if

$$\int_\Sigma (\alpha - \alpha') = \int_\Sigma d\beta = \int_{\partial\Sigma} \beta = 0. \tag{26.22}$$

Note, as before, it is important in this expression that β is defined throughout the manifold.

If we consider the structure of a massless chiral multiplet, we note that there are two scalars and a chiral fermion. In compactifications preserving $N = 1$ supersymmetry, modes of antisymmetric tensor fields which are annihilated by d will correspond to massless scalars; supersymmetry guarantees that the other elements of the multiplet are also present. The suggested readings at the end of the chapter contain more detailed discussion of these issues, but it is not too hard to understand how the various states in terms of the forms annihilated by d. The other massless *scalar* arises because one can also choose the form so the Laplacian vanishes. The Dirac operator is closely related to differential forms on manifolds. This can be shown using the creation–annihilation operator construction of the Dirac matrices we have used in our discussion of orthogonal groups. One can exhibit in this way the required pairing.

With this machinery, we can define an important set of topological invariants of manifolds: characteristic classes. Consider a gauge field, F, $F = dA$. Note that F is closed: $dF = 0$. Also, F is said to be an element of $H_1(M, R)$, the second

cohomology group of the manifold M with real coefficients. The cohomology class of such two forms is known as the first Chern class.

If the manifold is topologically non-trivial, then if we consider a gauge field, it may not be possible to describe the field everywhere by a single, non-singular potential. This problem is familiar to us from the case of the Dirac monopole. Instead, in different regions, α and β, we have to use different potentials. $A_{(\alpha)}$, $A_{(\beta)}$. In regions where α and β overlap (transition regions), $A_{(\alpha)}$ and $A_{(\beta)}$ will be gauge transforms of one another,

$$A_{(\alpha)} = A_{(\beta)} + \phi_{(\alpha\beta)}. \tag{26.23}$$

Another set of gauge fields are said to be in the same topological class if

$$\tilde{A}_{(\alpha)} = \tilde{A}_{(\beta)} + \phi_{(\alpha\beta)} \tag{26.24}$$

with the *same* transition function, ϕ. Now since the functions A and $\tilde{A}$ are not uniquely defined everywhere, $F = dA$ and $\tilde{F} = d\tilde{A}$ are not in the trivial cohomology class, in general. On the other hand, $F - \tilde{F}$ is, since the difference, $A - \tilde{A} = B$ is well defined. So $F - \tilde{F} = dB$, and F and $\tilde{F}$ are in the same cohomology class. So the cohomology class of F, the first Chern class, is a topological invariant.

There is a theorem that if the first Chern class is non-zero, one can always find a two-dimensional surface, Σ, with the property:

$$I(\Sigma) = \frac{1}{2\pi}\int_{\Sigma} F \neq 0. \tag{26.25}$$

Note that this is a kind of magnetic flux. By Dirac's argument, $I(\Sigma)$ is an integer. The first Chern class plays an important role in the theory of Calabi–Yau spaces.

These ideas can be generalized to complex spaces. Here we define, as we did for the orbifold, complex coordinates, z_i and $\bar{z}_i$. We then define a (p, q)-form ψ to be an object with p z_i-type indices and q $\bar{z}_i$-type indices. Note that ψ is totally antisymmetric in both types of indices. We can define two types of exterior derivatives, ∂ and $\bar{\partial}$, in the obvious way:

$$\partial\psi_{a_1\ldots a_{p+1},\bar{a}_1\ldots\bar{a}_q} = \frac{1}{p+1}\partial_{a_1}\psi_{a_2\ldots a_{p+2}\bar{a}_1\ldots\bar{a}_q} + (-1)^P \text{ permutations}. \tag{26.26}$$

Note that $\partial^2 = 0$; $\bar{\partial}$ is defined similarly. In terms of these definitions,

$$d = \partial + \bar{\partial}. \tag{26.27}$$

These are known as the Dolbeault operators. We can then consider differential forms annihilated by these operators. The numbers of independent forms annihilated by the ∂ and $\bar{\partial}$ operators are known as the Hodge numbers, $h^{p,q}$. Then, for example,

one has the Hodge decomposition:

$$b_n = \sum_{p+q=n} h^{p,q}. \tag{26.28}$$

Again, is is possible to choose these forms so that they are annihilated by the Laplacian.

26.2 Calabi–Yau spaces: constructions

We have already constructed a rather rich set of four-dimensional string theories. But they are only a small subset of what appears to be a vast set of possibilities. We saw, for example, that the orbifold compactifications give rise to moduli which describe states which are *not* orbifolds. A rich set of compactifications of string theory, of which the orbifolds we studied in the last chapter are special cases, are provided by the Calabi–Yau spaces. In this section, we introduce these.

Our strategy to construct solutions is to look for solutions of the ten-dimensional field equations. One can ask: why is this sensible? There are two answers. First, if we consider spaces in which the massless, ten-dimensional fields are slowly varying, it should be appropriate to integrate out the massive string modes and study the low-energy equations. A more serious question is: why can we simply look at the low-order equations? Even at the classical level, integrating out the massive states will lead to terms with arbitrary numbers of derivatives. This question is far more serious. If we solve the equations, say, involving two derivatives, we can try to find solutions of the terms up to four derivatives perturbatively. To do this, we expand the fields in modes of the lowest-order theory (e.g. eigenfunctions of the Laplace operator on the complex space). These are precisely the Kaluza–Klein modes. Calling these ϕ_n, plugging our lowest-order solution into the next order terms, we will obtain equations of the form:

$$\left(\nabla^2 + m_n^2\right)\phi_n = \frac{\alpha'}{R^2}\Gamma_n. \tag{26.29}$$

For $m_n \neq 0$, i.e. for the massive Kaluza–Klein modes, we simply obtain a small shift. But the massless modes are problematic. In the case of Calabi–Yau compactifications, it is supersymmetry which will come to our rescue. We will see that, for the massless modes, the tadpoles (Γ_ns) vanish.

We begin with the Type II theory. Rather than examine the equations of motion, we look at the supersymmetry variations. In flat space four-dimensional theories, we are familiar with the idea that we find minima of the potential by setting the auxiliary fields to zero. We can phrase this in a different, seemingly more obscure way. We can find static solutions of the classical equations by requiring that the

supersymmetry variations of all of the fields vanish. That is, we require:

$$\delta\psi = \epsilon F = 0 \quad \delta\lambda = \epsilon D = 0. \tag{26.30}$$

We will try the same strategy. In Chapter 17, we introduced the essential elements required to understand spinors in a gravitational background (the reader may want to reread Section 17.6). To make things simple, we will look for solutions where the antisymmetric tensor vanishes and the dilaton is constant, so only the metric is spatially varying. Then the condition that there be a conserved supersymmetry becomes:

$$\delta\psi_M = D_M\eta = 0. \tag{26.31}$$

So η is covariantly constant. This means that under parallel transport around any closed curve, η returns to itself. As in gauge theories, the effect of parallel transport can be described in terms of Wilson lines, where now the Wilson line is written in terms of the spin connection, ω:

$$U = Pe^{i\oint \omega\cdot dx}. \tag{26.32}$$

The fact that η is unchanged under any such transformation greatly restricts the form of ω. To see how this works, consider that in the ten-dimensional Lorentz group, there is an $O(6)$ which acts on the compactified coordinates, as well as the four-dimensional Lorentz group acting on the Minkowski coordinates. The 16-component spinor in ten dimensions decomposes under these groups as

$$\eta = (4, 2) + (\bar{4}, 2^*). \tag{26.33}$$

By local Lorentz transformations, we can take the (4, 2) to have the form (suppressing the four-dimensional spinor index):

$$\eta = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \eta_0 \end{pmatrix}. \tag{26.34}$$

In order that this be invariant, we require that the spin connection lie in an $SU(3)$ subgroup of $O(6)$. The space is said to be a space of $SU(3)$ holonomy.

In general, ω is an $O(6)$ matrix. Restricting to $SU(3)$ is a strong constraint. Already $U(3)$ holonomy requires that the manifold be complex. We encountered this already in the orbifold case, where we introduced three complex coordinates and their conjugates. There is no unique way to introduce the complex coordinates. The continuous set of choices will lead to a set of moduli of our solutions, known as the "complex structure moduli." In addition, a manifold of $U(3)$ holonomy is Kahler. This means that the metric can be derived from a function $K(x^i, x^{\bar{i}})$, the

Kahler potential, through:

$$g_{i\bar{j}} = \partial_i \partial_{\bar{j}} K. \tag{26.35}$$

While proving that a manifold of $U(3)$ holonomy must be Kahler is challenging, it is not hard to check that a Kahler manifold has $U(3)$ holonomy. Some aspects of these manifolds are discussed in the exercises.

The Christoffel symbols and curvature for a Kahler manifold can be written in quite compact forms. (Verification of these formulas is left for the exercises.) The components of the affine connection (Christoffel symbols) are given by

$$\Gamma^a_{bc} = g^{a\bar{d}} \partial_b g_{c\bar{d}} \quad \Gamma^{\bar{a}}_{\bar{b}\bar{c}} = g^{\bar{a}d} \partial_{\bar{b}} \partial_{\bar{b}} g_{\bar{c}d}. \tag{26.36}$$

As a result, the non-zero components of the Riemann tensor are:

$$R^{\bar{a}}_{\bar{b}c\bar{d}} = \partial_c \Gamma^{\bar{a}}_{\bar{b}\bar{d}} \tag{26.37}$$

and the Ricci tensor is

$$R_{\bar{b}c} = -\partial_c \Gamma^{\bar{a}}_{\bar{b}\bar{a}}. \tag{26.38}$$

Using

$$\Gamma^{\bar{a}}_{\bar{b}\bar{a}} = \partial_{\bar{b}} \ln \det g, \tag{26.39}$$

this can be further simplified:

$$R_{\bar{b}c} = -\partial_{\bar{b}} \partial_c \ln \det g. \tag{26.40}$$

Note that our result, Eq. (24.17), for the curvature of a two-dimensional Riemann surface is a special case of this.

The requirement that the metric have $SU(3)$ holonomy has a dramatic consequence for the curvature: the Ricci tensor vanishes. This follows from our discussion of the spin connection as a gauge field for local Lorentz transformations. On a six (real)-dimensional Kahler manifold, we have seen that the spin connection is not an $O(6)$ field, but rather a $U(3)$ field (in four dimensions, it is a $U(2)$ field, etc.). The $U(1)$ part of the Riemann tensor is the trace over the Lorentz indices – the group indices, thinking of the Riemann tensor as a non-Abelian field strength. But this object is the Ricci tensor. So $SU(3)$ holonomy requires that the Ricci tensor itself vanish everywhere on the manifold. For such a configuration, the lowest-order Einstein equation is automatically satisfied, $R_{i\bar{j}} = 0$. The question which we would like to address is, given a Kahler manifold, is it possible to deform the Kahler potential so that the Ricci tensor vanishes. Clearly a necessary condition for this is that the integral,

$$c_1 = \frac{1}{2\pi} \int \mathrm{Tr} R \tag{26.41}$$

vanish. This quantity is the first Chern class, the topological invariant which we discussed earlier. It was Calabi who conjectured that the vanishing of the first Chern class for a manifold was a necessary and sufficient condition that the manifold admit a unique metric of $SU(3)$ holonomy. Yau later proved this conjecture. The spaces constructed in this way are the famous Calabi–Yau spaces. In general, while one can prove that such metrics exist, actually constructing them is a difficult numerical problem. Fortunately, many properties relevant to the low-energy behavior of string theory on these manifolds can be obtained from more limited, topological information.

It is worthwhile comparing this with our orbifold constructions. The orbifolds are everywhere flat. But the existence of a deficit angle associated with the fixed points means that there is actually a δ-function curvature; this gives precisely the holonomy of these manifolds. If we decompose the spinors, as before, then as we transport them about the fixed points, the i components pick up a phase, $e^{\frac{2\pi i}{3}}$, while the 0 components are invariant. Correspondingly, we found one unbroken supersymmetry.

When we discuss the heterotic theory on a Calabi–Yau space, we will have to choose values for the gauge fields as well. It will not be possible to simply set the gauge fields to zero. From the point of view of four dimensions, gauge fields with indices in the extra dimensions are like scalars, so this will result in breaking of some or all of the gauge symmetry. As we will see in Section 26.6.1, there are many possible choices for these fields, with distinct consequences for the structure of the low-energy theory. In an interesting subclass, some features of the heterotic theory are closely related to those of Type II on Calabi–Yau spaces.

26.3 The spectrum of Calabi–Yau compactifications

In both the Type II and heterotic cases, many features of the low-energy spectrum follow from general topological features of the manifold and do not depend on details of the metric. In the heterotic case, the number of generations (minus the number of anti-generations) is a topological invariant. Suppose that we have some number of generations for some choice of metric. If we now make smooth, continuous changes in the metric, the massless spectrum can change as generations and anti-generations pair to gain mass or become massless. In other words, a mass term in an effective action can pass through zero, but the net number of generations cannot change. In some cases, other features of the spectrum are similarly invariant. So while it is difficult to write down explicit metrics for manifolds of $SU(3)$ holonomy, it is possible to determine many important features of the low-energy theory from basic topological features of the manifold.

In the Type II theory, the numbers of hypermultiplets and vector multiplets are separately topological. They do not pair up as one moves about on the moduli space; the $N = 2$ supersymmetry insures that if a field is massless at one point in the moduli space, it is massless at all points. Even more dramatic is the fact that massless states found in the lowest order of the α' expansion are exactly massless. So it is enough to study the lowest-order supergravity equations of motion to count massless particles.

The important non-zero Hodge numbers are $h^{2,1}$ and $h^{1,1}$. In the IIA theory, there are $h^{1,1}$ vector multiplets and $h^{2,1}$ hyper-multiplets. In the IIB theory, this is reversed. In the heterotic case, the (2, 1) forms will correspond, effectively to generations, the (1, 1)-forms to anti-generations.

The counting of massless fields is not difficult to understand. Since we have taken the antisymmetric tensor fields and fermions to vanish in the background, the equations for these fields are particularly simple. Consider the antisymmetric tensor, $B_{\mu\nu}$. On a complex manifold, as we explained earlier, there are $h^{1,1}$ (1, 1) forms, $b^{(a)}_{i,\bar{j}}$ and $h^{2,1}$ (2, 1)-forms annihilated by the operators ∂ and $\bar{\partial}$. Since the corresponding three-index field strengths, $H = dB$, vanish, there is no energy cost to giving a constant expectation value to the associated four-dimensional fields; they correspond to massless scalars in four dimensions. The fields connected to the (1, 1)-forms, $b_{i\bar{i}}$, are easy to describe. In addition to the antisymmetric tensor, there is also a massless perturbation of the metric:

$$ig_{i\bar{i}}(x, y) = \phi(x) b_{i\bar{i}}(y). \tag{26.42}$$

Here x refers to the ordinary four-dimensional Minkowski coordinates, and y refers to the compactified coordinates. Similarly, in the IIA theory, one can find a massless gauge field, rounding out the bosonic components of the vector multiplet. This comes from the three-index Ramond field,

$$C_{\mu i\bar{i}}(x, y) = A_\mu(x) b_{i\bar{i}}(y). \tag{26.43}$$

We will leave for the reader the problem of working out the structure of the hypermultiplets in terms of the (2, 1)-forms, and also determining the pairings in the IIB case.

One (1, 1)-form which is always present is the Kahler form:

$$b^K_{i\bar{i}} = ig_{i\bar{i}};\ b_{\bar{i}i} = -ig_{i\bar{i}}. \tag{26.44}$$

This satisfies

$$\partial b^K = \bar{\partial} b^K = 0 \tag{26.45}$$

because $g_{i\bar{i}} = \partial_i \partial_{\bar{i}} K$. The real scalar which sits in the multiplet with b^K is just the metric itself. The corresponding massless field is the radius of the compact

space:

$$g_{i,\bar{i}}(x^\mu, z^i) = R^2(x^\mu) g_{i,\bar{i}}(z); \quad B_{i,\bar{i}}(x^\mu, z^i) = b(x^\mu) b_{i\bar{i}}(z). \tag{26.46}$$

That the field is massless is no surprise; the condition $R_{i\bar{i}} = 0$ is not changed under an overall rescaling of the metric so the vev is undetermined.

26.4 World sheet description of Calabi–Yau compactification

So far, we have described compactification of string theory in terms of ten-dimensional space-time. This analysis makes sense if the radius of the compactified space is large compared to the string length, ℓ_s. We can also formulate these questions in world sheet terms. This provides a complementary way to understand many features of the compactified theory. This is useful for at least two reasons. First, it provides tools to ask what happens when the compactification radius is of order the string scale or smaller. Second, there are some features of the spectrum and interactions which are more readily accessible in this framework.

In the Type II theory, the non-linear sigma model which describes compactification on a Calabi–Yau space has some striking features. First, in the absence of background antisymmetric tensor fields, it is left–right symmetric. Second, there are two left-moving and two right-moving supersymmetries on the world sheet, as opposed to the one left-moving and one right-moving supersymmetry of a general configuration. This can be usefully understood in a number of ways. In the light cone gauge, one can work with the covariantly constant spinor η and its conjugate, $\bar{\eta}$, to construct two left-moving and two right-moving supersymmetry generators, both in the sense of the world sheet and in space-time. We have already seen this in the case of orbifold constructions. There, in the light cone gauge, we have eight left-moving and eight right-moving supersymmetry generators, before the orbifold projection. We can organize these in terms of their transformation properties under the $SU(3) \times U(1)$ holonomy group. For both the left and right movers, there are triplets, Q_i, anti-triplets, $\bar{Q}_{\bar{i}}$, and singlets, Q_0 and $\bar{Q}_0$. The triplets and anti-triplets are charged under the $U(1)$; the singlets are not. The orbifold projection eliminates the triplets. The two singlets survive.

In a purely world sheet description, non-linear sigma models described by a Kahler metric automatically have two left-moving and two right-moving supersymmetries. To describe these, we can introduce a superspace with four Grassmann coordinates, two left movers and two right movers: θ_+^A and θ_-^A. This superspace can be thought of as the truncation of $N = 1$ supersymmetry in four dimensions. We can define, as in four dimensions, operators D_α and $\bar{d}_\alpha$, and left- and right-moving chiral fields annihilated by the $\bar{d}$s. Correspondingly, we can define chiral left- and

right-moving fields:

$$X^i_+(z,\theta) = x^i(z) + \theta^A_+ \psi^i_A(z) + \text{auxiliary field} \tag{26.47}$$

and similarly for $X^{\bar{i}}_-$. In terms of these fields, we can write the action of the conformal field theory as:

$$\int d^2\sigma \int d^2\theta_+ d^2\theta_- K(X, \bar{X}). \tag{26.48}$$

Integrating over the θs, the bosonic terms are just $\int d^2\sigma\, g_{i\bar{i}} \partial_\alpha x^i \partial^\alpha x^{\bar{i}}$, with $g_{i\bar{i}}$ the Kahler metric.

The superconformal algebra, in these backgrounds, is enlarged to what is referred to as the $N = 2$ superconformal algebra (one such algebra for the left movers, one for the right movers). In addition to the stress tensor and the two supercurrents, this algebra contains a $U(1)$ current. The supersymmetry generators can be constructed by the Noether procedure. They can also be guessed by taking the generators in a flat background, and making the expressions covariant:

$$G^+ = g_{i\bar{i}} DX^i \psi^{\bar{i}} \quad G^- = g_{i\bar{i}} D X^{\bar{i}} \psi^i. \tag{26.49}$$

These have opposite charge under the $U(1)$ current (an R current) constructed from the fermions:

$$j(z) = \psi^{\bar{i}}(z)\psi^i(z) \tag{26.50}$$

and a similar current for the left movers. The full algebra is:

$$\begin{aligned} T(z)G^\pm(0) &\approx \frac{3}{2z^2} G^\pm(0) + \frac{1}{z}\partial G^\pm(0) \\ T(z)j(0) &\approx \frac{1}{z^2} j(0) + \frac{1}{z}\partial j(0) \\ j(z)G^\pm(0) &\approx \pm\frac{1}{z} G^\pm(0). \end{aligned} \tag{26.51}$$

These equations say that G has dimension $3/2$, while j has dimension 1, and $G^\pm$ have $U(1)$ charges plus and minus one. The central charge appears in the relations:

$$\begin{aligned} G^+(z)G^-(0) &\approx \frac{2c}{3z^3} + \frac{2}{z^2} j(0) + \frac{2}{b} T(0) + \frac{1}{z}\partial j(0) \\ G^+(z)G^+(0) &\approx 0 \\ j(z)j(0) &\approx \frac{c}{3z^2}. \end{aligned} \tag{26.52}$$

The non-linear sigma models appropriate to *heterotic* compactifications on Calabi–Yau spaces have a number of interesting features. We will see that, for a particular choice of gauge fields, the world sheet theory which describes the heterotic

compactification is identical to that of the Type II theory. Thus, again, they have two left-moving and two right-moving supersymmetries ((2, 2) supersymmetry). The fact that the world sheet theories of the two different string theories are the same allows us to argue, as we will below, that Calabi–Yau spaces are solutions of the full, non-perturbative string equations of motion. But this observation also tells us about interesting features of the spectrum.

To understand the spectrum, it is helpful to ask, first, what is a vertex operator from the perspective of the two-dimensional conformal field theory? The answer is that a vertex operator is a *marginal deformation* of the theory, a perturbation of dimension 2 ((1, 1), in terms of the left- and right-moving Virasoro algebras). The standard way to compute the dimensions of operators is to treat them as perturbations, and calculate, for example, the beta function of the perturbation. For marginal operators, the beta function vanishes to first order. Moduli correspond to "exactly marginal deformations" of the theory. For these, the beta functions vanish to all orders in the perturbation (and non-perturbatively), corresponding to the fact that the theory, even for a finite perturbation, is conformal.

The existence of moduli means that there is a multiparameter set of conformal field theories. Varying the action with respect to the parameters yields operators which are exactly marginal. In this way, we have the two-dimensional version of the correspondence between moduli and massless fields.

An example of a modulus is the radius of the complex space. The lowest-order equation for the metric is invariant under an overall scaling of lengths. But this is not obviously true of the higher-order corrections. For Type II theories, the space-time supersymmetry guarantees that there is no potential for the moduli, so the sigma model is a good conformal field theory, suitable for heterotic string compactification. On the heterotic side, we can also give a more direct world-sheet argument. Here R^{-2} is the coupling constant of the σ model. In other words, writing the metric as R^2 times a reference metric of order the string scale, R^2 scales out of the Lagrangian. We know that the lowest-order beta function equation is the same as the field theory equation. It is trivially independent of R^2, since it is a one-loop effect. For higher orders, there is a non-renormalization theorem. This follows from a combined world sheet, space-time argument. The superpartner of fluctuations in the radius is the fluctuation of the antisymmetric tensor field, $b_{i\bar{i}} = ig_{i\bar{i}}$. The associated vertex operator (term in the action) is a total derivative on the world sheet at zero momentum. It is perhaps easiest to see this by writing the vertex operator at zero momentum in the form:

$$\begin{aligned} V_b &= b_{MN}\epsilon^{\alpha\beta}\partial_\alpha X^M \partial_\beta X^N \\ &= \partial_M \partial_N K \epsilon^{\alpha\beta}\partial_\alpha X^M \partial_\beta X^N \\ &= \partial_\alpha[\epsilon^{\alpha\beta}\partial_\beta X^M \partial_M K]. \end{aligned} \tag{26.53}$$

So b decouples at zero momentum. Because b is in a supermultiplet with R^2, this means that the superpotential, which is a holomorphic function of the superfields, is independent of R^2.

Actually, this statement is not precisely correct, because K is not single-valued. In perturbation theory, it is true, since one is not sensitive to the global structure of the manifold (in perturbation theory, all fluctuations are small). Non-perturbatively, one can encounter instantons in the world sheet theory. A more detailed analysis is required to determine whether there are corrections to the superpotential. In left–right symmetric compactifications of the heterotic string, those with two left-moving and two right-moving supersymmetries ((2, 2) models), a study of fermion zero modes in the presence of the instanton shows that no superpotential for the moduli is generated; this is consistent with the expectations from the Type II theory. For compactifications with two right-moving but no left-moving supersymmetries ((2, 0) models), corrections can be generated, though in some cases intricate cancellations still prevent the appearance of a potential for the moduli. These two classes of models are phenomenologically quite distinct, as we will see shortly.

26.5 An example: the quintic in CP^4

It is helpful to have a concrete example of a Kahler manifold with $c_1 = 0$, on which we know one can construct a metric of $SU(3)$ holonomy. We have previously encountered the complex projective spaces in N dimensions, CP^N. These are defined as the spaces with $N + 1$ complex coordinates, Z_a, with the identification $Z_a \to \lambda Z_a$, for any complex number λ. We have written down a Kahler potential on this space:

$$K = \ln\left(1 + \sum_{a=1}^{N} Z_a \bar{Z}_a\right). \tag{26.54}$$

Any complex submanifold of a Kahler manifold is also a Kahler manifold; one can simply take the Kahler potential to be the Kahler potential of the full manifold evaluated on the submanifold. To obtain a manifold with three complex dimensions, we can start with CP^4, and write an equation for the vanishing of a polynomial, $P(Z)$. The polynomial should be homogeneous, in order that it have a sensible action in CP^N. It turns out that it should satisfy other conditions. Its gradient should vanish, at most, at the origin (which is not a point in CP^N). In order that the first Chern class vanish, it should be quintic. We will give an argument for this shortly.

The simplest (most symmetric) possibility is:

$$P = Z_1^5 + Z_2^5 + Z_3^5 + Z_4^5 + Z_5^5 = 0, \tag{26.55}$$

but there are obviously many more. We can deform this polynomial by adding other quintic polynomials. These correspond to varying the complex structure. Since each deformation produces another solution of the string equations, each deformation corresponds to a modulus, the "complex structure moduli." Associated with each of these deformations is a form of type (2, 1), which we will not attempt to construct here.

Before listing the deformations, note that not every deformation corresponds to a change of the physical situation – and thus to a massless particle. Holomorphic changes of the coordinates which are non-singular and invertible do not change the complex structure. The transformation

$$Z_i \rightarrow Z_i + \epsilon^{ij} Z_j \tag{26.56}$$

is well-defined in CP^4. As a consequence, deformations such as $Z_1^4 Z_2$ are not physical. So we can list the possible deformations:

$$Z_1^3 Z_2^2 \ldots;\ Z_1^3 Z_2 Z_3 \ldots;\ Z_1^2 Z_2^2 Z_3, \ldots;\ Z_1^2 Z_2 Z_3 Z_4, \ldots;\ Z_1 Z_2 Z_3 Z_4 Z_5. \tag{26.57}$$

All together there are 101 possible deformations of the polynomial, corresponding to $h_{2,1} = 101$. In this example, there is only one Kahler modulus, the overall radius of the compact space.

We can understand heuristically why the first Chern class vanishes in a way which we will help us understand other features of these manifolds. A characteristic feature of the Calabi–Yau spaces is the existence of a covariantly constant 3-form, ω_{ijk}. The existence of this form follows from the existence of the covariantly constant spinor, η:

$$\omega_{ijk} = \bar{\eta} \Gamma_{[ijk]} \eta. \tag{26.58}$$

Working in terms of the creation–annihilation operator basis for the Γs, one sees that ω is holomorphic. The Γ_is can be defined so that the $\Gamma_{\bar{i}}$ matrices annihilate η. Then, because of the complete antisymmetrization, only components of ω with i, j, k indices are non-vanishing. In the space defined by the vanishing of a quintic polynomial in CP^4, we can show that there exists a holomorphic three-form which is everywhere non-vanishing. Calling $x^i = Z_i / Z_5$, $i = 1, \ldots, 4$:

$$\omega = dx^1 \wedge dx^2 \wedge dx^3 \left(\frac{\partial P}{\partial x^4} \right)^{-1}. \tag{26.59}$$

One can show that this expression does not depend on singling out a particular coordinate, and that it is not singular at the points where the derivative vanishes, provided that the polynomial P is quintic and that the gradient of P vanishes only at the origin. The existence of such a form can be shown to be equivalent to the vanishing of the first Chern class.

26.6 Calabi–Yau compactification of the heterotic string at weak coupling

Much effort has been devoted to the study of compactifications of the weakly coupled heterotic string on Calabi–Yau spaces. These theories have many of the features of the Standard Model. They also allow one to consider many of the questions of beyond the Standard Model physics. Before beginning an analysis of these models, it is worth listing some of the points that we can address in this framework.

(1) Low-energy supersymmetry: solutions of the classical equations of the heterotic string theory on Calabi–Yau spaces exist. They have $N = 1$ supersymmetry. Supersymmetry, as in field theory, is unbroken to all orders of perturbation theory, but may be broken non-perturbatively.
(2) Low-energy gauge groups: the simplest constructions have gauge group $E_8 \times E_6$, broken perhaps by Wilson lines, which preserve the rank of the gauge group. But many models have a moduli space in which the gauge group is broken to precisely that of the Standard Model.
(3) Generations: the number of generations is typically determined in terms of topological features of the underlying manifold.
(4) Massless particles, not protected by symmetries: various massless states arise, which are not protected by chiral symmetries. This is precisely what we want in order to understand the presence of light Higgs fields in supersymmetric theories. We know that if such fields are present in the low-energy field theory, they are protected from gaining large masses by non-renormalization theorems. In field theory, the vanishing of such mass terms appears mysterious; in these string constructions, it is automatic. Such states could play the role of Higgs fields in supersymmetric models. In other words, *the μ problem of ordinary supersymmetric field theories is readily solved in this framework.*
(5) Unification of couplings: the string theories we are studying are not grand unified theories in the conventional sense. There is no energy scale at which these compactifications appear as four-dimensional theories with a single unbroken gauge group. Yet, generically, couplings are unified. These last two points, which we will see are easy to understand in terms of the microscopic structure of string theory, are quite surprising from a low-energy point of view. They have sometimes been referred to as "string miracles."
(6) Continuous and discrete symmetries: it is easy to prove that, for these compactifications (and for weak-coupling heterotic models in general) there are no continuous global symmetries; all continuous symmetries must be gauge symmetries. Discrete symmetries, on the other hand, proliferate, and might play the role of R-parity or lead to other interesting phenomena. These discrete symmetries are typically gauge symmetries, in the sense that they are residual symmetries left over after the breaking of continuous gauge symmetries.

We will also see that there are a number of problems with these models, which illustrate some of the basic difficulties in developing a string phenomenology.

(1) There are too many of them. While there are models with three generations (many), there are models with hundreds of generations, with non-standard gauge groups and the like.
(2) The problem of moduli: non-perturbatively, moduli can acquire potentials, but they typically vanish in various asymptotic regimes. Simple general arguments indicate that stable, supersymmetry-breaking minima, if they exist, must be in regions which are inherently strongly coupled, in the sense that *no* weak coupling approximation is available.
(3) The problem of the cosmological constant: this is closely related to the previous one. In many instances moduli potentials can be calculated. For any given value of the moduli, the size of these potentials is scaled as one would expect by the scale of supersymmetry breaking. As a result, even if strongly coupled, stable minima exist, it is not clear why the cosmological constant should be small at these points.

We will not offer a solution to these problems in this chapter, but will explore at least one proposed answer known as the "landscape" in the concluding chapter.

26.6.1 Features of Calabi–Yau compactifications of the heterotic string

In the previous section, we asserted that, in suitable backgrounds, the world sheet conformal field theory which describes the heterotic string is the same as that which describes the Type II theory. Here, we describe compactifications of the heterotic string theory in more detail.

To construct solutions, we still look for solutions which preserve a space-time supersymmetry. Again, we require the supersymmetry variation of the gravitino to vanish, giving $D_\mu \eta = 0$, so once more we need a covariantly constant spinor. There is now an equation for the variation of the ten-dimensional gaugino, as well:

$$\delta\lambda \propto \Gamma^{ij} F_{ij} \eta. \tag{26.60}$$

One strategy, then, to find solutions which preserve $N = 1$ supersymmetry is to require that $F_{ij}\Gamma^{ij}$ is an $SU(3)$ matrix. There is a simple Ansatz which achieves this. Both E_8 and $O(32)$ have $SU(3)$ subgroups:

$$SU(3) \times E_6 \times E_8 \subset E_8 \times E_8 \quad SU(3) \times O(26) \subset O(32). \tag{26.61}$$

On the Calabi–Yau space, the spin connection is an $SU(3)$ valued field, so take the gauge field to be a field in one of these $SU(3)$ subgroups. Then for gauge generators not in $SU(3)$, Eq. (26.60) is automatically satisfied. For those in $SU(3)$, the condition is mathematically identical to that for the gravitinos, and is again satisfied.

This Ansatz satisfies another condition. We put the antisymmetric tensor field B to zero, but, because of the Chern–Simons terms, this does not by itself guarantee that the field strength H is zero. But with this Ansatz, the Chern–Simons terms for

the gauge and gravitational fields are identical. As a quick check, note that

$$dH = \mathrm{Tr}\, R \wedge R - \mathrm{Tr}\, F \wedge F, \tag{26.62}$$

and these terms clearly cancel.

This establishes that this is a solution of the equations of motion to lowest order in the α' expansion. But there is another way to see this, which will allow us to establish, as we did for the Type II theory, that this is an *exact* solution, perturbatively and non-perturbatively. If we write the non-linear sigma model which describes the heterotic string in this background, it is identical to that for the Type II theory. To see this, as in the orbifold case, we divide the left-moving gauge fermions into three sets. First, there are the fermions in the "other" E_8, which are not affected by the background gauge field and remain free, λ^A, $A = 1, \ldots 16$. In the first E_8, 10 fermions, λ^a, $a = 1, \ldots, 10$ (transforming as a vector in the $O(10)$ subgroup of E_6) are also free. The remaining six interacting fermions can be grouped, like the left-moving coordinates, into three complex fermions, λ^i and $\lambda^{\bar{i}}$. These fermions interact in precisely the same way as the left-moving fermions in the Type II theory. This can be seen by writing the action of the Type II fermions in terms of the vierbein and spin structure, rather than the metric and the Christoffel connection.

We see from this that the moduli of the Type II theory are also moduli of the heterotic theory. Actually, we knew this had to be, since we know that each of these conformal field theories, on the Type II side, is a good conformal field theory for the heterotic theory. But we can also see this pairing more directly in the language of vertex operators. Here it is somewhat more convenient to work in the RNS picture. The vertex operators correspond to small deformations of the background in the directions of the moduli. In the Type II theory, they are built from right-moving fields, ∂X^i and ψ^i, and left-moving fields, $\bar{\partial} X^i$, $\tilde{\psi}^i$. In the heterotic case we can trade $\tilde{\psi}^i$ with λ^i. Since the action for the λ^is is the same as for the $\tilde{\psi}^i$s, the dimensions of the vertex operators are exactly the same. This does not preclude the existence of additional moduli on the heterotic side, and we will see that there *are* typically additional moduli in these compactifications.

While all moduli of the Type II theory are moduli of the heterotic theory, not all heterotic moduli correspond to states of the Type II theory. Vertex operators for moduli which preserve only two right-moving supersymmetries ((2, 0)) are not suitable vertex operators for the Type II theory. The moduli we are considering here are distinguished because they preserve the two left-moving world sheet supersymmetries, and we will refer to these as Type II moduli. Perhaps more interesting, though, than the pairing of moduli is is a pairing of the Type II moduli with matter fields. The moduli associated with (2, 1)-forms are paired with 27s of E_6; (1, 1) moduli with $\overline{27}$s. This is most readily seen in the language of vertex operators, using the world sheet superconformal symmetry. The vertex operators for the Type II theory are the highest components of the corresponding superconformal multiplets with

respect to both left- and right-moving supersymmetries. In superspace, they are the $\theta_+^2\theta_-^2$ components of operators of the form:

$$f(X^i, \bar{X}^i). \tag{26.63}$$

The $\theta_+\theta_-^2$ component has dimension $(1/2, 1)$. We can form an operator of dimension $(1, 1)$ by multiplying by λ^a, one of the free fermions. This operator is not highest weight with respect to the left-moving $N = 2$ algebra, but this is not a problem; this symmetry is not a gauge symmetry on the world sheet, but simply an "accident" of our choice of background field. It is highest weight with respect to the left-moving Virasoro algebra, which is all that matters.

We already observed this pairing in the Z_3 orbifold model, which is a special case of the Calabi–Yau construction. In the untwisted sector, the vertex operators for the moduli took the form, on the left:

$$\bar{\partial} X^i, \tag{26.64}$$

while for the 27s they took the form:

$$\lambda^a \lambda^i. \tag{26.65}$$

The supersymmetry transformation of the latter operator changes λ^i to $\bar{\partial} X^i$.

The distinction between 27s and $\overline{27}$s is readily understood. In the Type II case, we can distinguish two types of moduli, depending on their charges under the $U(1)$ within the left-moving $N = 2$ algebra. In the orbifold context some vertex operators involve $\bar{\partial} X^i$, some $\bar{\partial} X^{\bar{i}}$. In the heterotic case, the world sheet $U(1)$ symmetry corresponds to the $U(1)$ subgroup of E_6 in the decomposition $O(10) \times U(1) \subset E_6$. This $U(1)$ charge is precisely what distinguishes the 10s, for example, in the 27 and $\overline{27}$. In the Type II case, this distinction corresponds to the distinction between $(2, 1)$ and $(1, 1)$ moduli, so we obtain precisely the pairing we described above (note that what one calls a 27 and a $\overline{27}$ is a matter of convention; if one adopts the opposite convention, the identification is reversed).

This result holds everywhere in the moduli space; since the number of moduli of each type does not change as one moves in the moduli space, the number of 27s and $\overline{27}$s does not change. This is a surprising result. One might have thought that, in a complicated construction such as this, 27 and $\overline{27}$s would, whenever possible, pair to gain mass. But this is not the case. This is precisely the sort of phenomena one needs to understand light Higgs particles in supersymmetric theories. We will see shortly how this works in a more detailed model.

26.6.2 Gauge groups: symmetry breaking

So far, the heterotic models we are considering have group $E_8 \times E_6$. If we are to describe the Standard Model, we need to be able to break this symmetry. We have

seen in the case of toroidal compactifications that gauge symmetries can be broken by expectation values of gauge fields with indices in compactified dimensions. Stated in a more gauge-invariant fashion, these are non-trivial expectation values for Wilson lines. In the Calabi–Yau case, the same is possible.

Let's consider a specific example: the quintic in CP^4, with vanishing of the polynomial:

$$Z_1^5 + Z_2^5 + Z_3^5 + Z_4^5 + Z_5^5 = 0. \tag{26.66}$$

The corresponding Calabi–Yau manifold, we saw, has 101 27s and one $\overline{27}$. This polynomial has a variety of symmetries. As for the torus, we can use these to project out states and simplify the spectrum. Consider, for example, the symmetry:

$$Z_i \to \alpha^i Z_i \quad \alpha = e^{\frac{2\pi i}{5}}. \tag{26.67}$$

This is a symmetry of the polynomial. It is somewhat different than the orbifold symmetries we have discussed, since, as the reader can check it *acts without fixed points.* Mathematicians call such a symmetry "freely acting." For physics, it means that if we mod out the Calabi–Yau by this symmetry, while it is still necessary to include twisted sectors, the twisted strings have mass of order R, the Calabi–Yau radius, and there are no light states in these sectors if R is large.

We can readily classify the states invariant under this symmetry. Among the moduli, there are 21 $h_{2,1}$ fields, associated with polynomials such as $Z_1^3 Z_3 Z_4$ and $Z_1 Z_2 Z_3 Z_4 Z_5$. The Kahler modulus (overall radius) is also invariant under this transformation, and so survives the projection. The corresponding Euler number is 40, 1/5 the Euler number of the covering space. There are, as well, 21 27s of E_6 and one $\overline{27}$. Further symmetries can be used to reduce the number of generations to as few as four.

But what interests us here is obtaining smaller gauge groups. We can define the Z_5 to include a transformation in E_6. This is equivalent to the presence of a Wilson line on the manifold. An interesting way to do this is to consider a somewhat different decomposition of E_6 than we have considered up to now:

$$SU(3) \times SU(3) \times SU(3) \subset E_6. \tag{26.68}$$

An example of a Wilson line in this product of $SU(3)$s is:

$$U = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & \alpha^3 \end{pmatrix} \times \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & \alpha^3 \end{pmatrix}. \tag{26.69}$$

This breaks E_6 to $SU(3) \times SU(2) \times SU(2) \times U(1)^2$.

26.6.3 Massless Higgs fields, or the μ problem

When we mod out so as to reduce the gauge symmetry, we also alter the spectrum. We have seen that we greatly reduce the number of moduli and the number of generations. The presence of the Wilson lines also disrupts the left–right symmetry of the model. As a result, the pairing of moduli and matter fields is no longer quite so simple.

In the presence of the Wilson line, one still obtains 20 complete E_6 generations. If one thinks, loosely, of some of the massless fields "gaining" mass, elements of the 27 and $\overline{27}$s must pair up to gain mass. More precisely, in this modding out procedure, states disappear, but they must disappear in pairs. But one also obtains some incomplete multiplets, where paired states do not disappear. Consider the $\overline{27}$. This is invariant under the original Z_5s, so any state which survives must be invariant under the Wilson line. Using the decomposition of the 27 under $SU(3)^3$:

$$27 = (3, 1, \bar{3}) + (\bar{3}, 3, 1) + (1, \bar{3}, 3). \tag{26.70}$$

So we obtain Z_5 singlets from only the third multiplet. These form a $(1, 2, 2)$ under $SU(3) \times SU(2) \times SU(2)$, as well as a singlet. There is a corresponding pair of states from the 27s. This is the sort of multiplet we would like to understand the presence of light Higgs particles in supersymmetric models: massless states, at tree level, which arise, from a low-energy point of view, more or less by accident.

26.6.4 Continuous global symmetries

In the heterotic string theory, there are no continuous global symmetries. We won't give the formal proof here, but the basic argument is not hard to understand. If there is a global symmetry, it should be a symmetry of the world sheet theory. In this way, we are guaranteed that vertex operators can be chosen to have well defined transformation properties, and so the S-matrix will transform properly. The global symmetry will be associated with a world sheet current. This current can be decomposed into left- and right-moving pieces. But from any left-moving current, we can build a gauge boson vertex operator, so the symmetry is necessarily a gauge symmetry. Right-moving currents will not commute with the world sheet supersymmetry generators, and will not have well-defined action on states (in the BRST language, they do not commute with the BRST operator). So they are not symmetries in space-time.

This argument also indicates that there are no global symmetries in the Type II theories. This is in accord with our expectation that global symmetries are unlikely to arise in a theory of quantum gravity.

26.6.5 Discrete symmetries

When we studied orbifold models, we found discrete symmetries existed in a subset of vacua on the full moduli space. This is also the case for the Calabi–Yau manifold constructed from the vanishing of a quintic polynomial in CP^4. Such symmetries turn out to be quite common.

The quintic polynomial, $P = \sum Z_i^5$, exhibits a set of Z_5 symmetries:

$$Z_i \to \alpha^{k_i} Z_i \quad \alpha = e^{\frac{2\pi i}{5}}. \tag{26.71}$$

An overall phase rotation of all of the Z_is has no effect in CP^4, so the symmetry here is Z_5^4. There is also a permutation symmetry, S_5. This symmetry group is a subgroup of the $O(6)$ symmetry which would act on six non-compact, flat dimensions. We can thus think of these symmetries as discrete gauge transformations. So their existence in a theory of gravity is not surprising.

We would like to know if these symmetries are R-symmetries or not. We can address this by asking their effect on the covariantly constant spinor, η. This is more challenging to do than in the orbifold context, since we do not have quite such explicit expressions. It is simplest to look at the covariantly constant 3-form. We already gave a construction:

$$\omega = dx^1 \wedge dx^2 \wedge dx^3 \left(\frac{\partial P}{\partial x^4}\right)^{-1} \tag{26.72}$$

with $x^i = Z_i/Z_5$. This construction treats the coordinates asymmetrically, but, as we explained, ω is symmetric among the coordinates. Note that ω transforms essentially like η^2, i.e. like θ^2. So symmetries under which ω transforms non-trivially are R-symmetries, and W transforms like ω.

Consider, first, the Z_5 transformations of the separate Z_is. We can read off immediately how ω transforms under transformations of the first three; the other two follow by symmetry. So

$$\omega \to \alpha^{\sum k_i}. \tag{26.73}$$

Similarly, under those S_5 transformation which permute Z_1, Z_2, Z_3, we can see how ω transforms. If the permutation is odd, ω changes sign. So again, the general S_5 transformation is an R-symmetry.

Turning on the various complex structure moduli typically breaks some or all of this symmetry. For example, if we turn on the modulus associated with the polynomial

$$z_1 z_2 z_3 z_4 z_5 \tag{26.74}$$

we break the Z_5^4 symmetry down to a subgroup satisfying $\sum k_i = 0 \bmod 5$. This group is Z_5^3, but it is a non-R-symmetry, in light of the transformation law of ω. An expectation value for this field clearly preserves the permutation symmetry.

Similarly, turning on, say, $aZ_1^3Z_2 + bZ_1^2Z_2^3$ breaks the symmetries acting on Z_1 and Z_2, as well as some of the permutation symmetry. Turning on enough fields breaks all of the symmetry.

One might ask why one should be interested in points or surfaces in the moduli space which preserve a discrete symmetry, when in the bulk of the space there is no symmetry. This question is closely related to the question: what sorts of dynamics might fix the moduli? This is a subject which we will deal with extensively later, but for which we will provide no definitive resolution. But even without understanding this dynamics, there is a simple reason to suspect that points in the moduli space with symmetries might be singled out by dynamics. Imagine that we somehow manage to compute an effective potential for the moduli, arising, perhaps, due to some non-perturbative string effects. Symmetry points are necessarily *stationary points* of this effective potential. There is, of course, no guarantee that they are *minima* of the potential, but they are certainly of interest as candidates for string ground states.

There are, as we have seen, certain facts of nature which suggest that discrete symmetries might play some role in extensions of the Standard Model, including proton decay and dark matter.

26.6.6 Further symmetry breaking: the Standard Model gauge group

The Wilson line mechanism, as we have described it, provides a path to reduce the gauge symmetry from $E_6 \times E_8$, but leaves the rank untouched.[1] We can hope to reduce the gauge symmetry further by giving expectation values to some of the matter fields. Ideally, these expectation values would be large. The presence of other gauge groups (as well as unwanted matter multiplets) can spoil the prediction of coupling unification, and lead to severe difficulties with proton decay and other rare processes. We are led, then, to ask if we can consider more general states, in which the spin connection is not equal to the gauge connection, and the rank is reduced.

This is a complex subject, which has only been partially explored. At lowest order in the α' expansion, there are such flat directions. They are not left–right symmetric, and while, in order that they exhibit space-time supersymmetry, they have two right-moving supersymmetries, they have only one left-moving supersymmetry. So they are not suitable backgrounds for Type II theories, and one cannot argue as easily as

[1] Non-Abelian discrete symmetries offer possibilities for reducing the rank, but we will not explore these here.

for the standard embedding that these $(0, 2)$ configurations are solutions of exact classical string equations. They are still subject to perturbative non-renormalization theorems in α'. But detailed study of instanton amplitudes is required to determine if these flat directions are lifted non-perturbatively, i.e. by corrections of the form $e^{-R^2\alpha'}$.

There are, however, a class of vacua with Standard Model gauge group which can be found by symmetry arguments, much as we found additional flat directions in the Z_3 orbifold model. Consider, again, the quintic in CP^5, with the symmetric polynomial. We can find flat directions of the D terms by taking $27 = \overline{27}$. More precisely, starting with the E_6 decomposition into $O(10)$ representations,

$$27 = 10^1 + 1_{-2} + 16_{-1/2}, \tag{26.75}$$

we can give expectation values to the singlets in the $\overline{27}$ and one of the 27s. These are also flat directions of the F terms. For example, consider the 27 corresponding to the polynomial $Z_1Z_2Z_3Z_4Z_5$. The product, $\overline{27}\ 27$ is invariant under all of the discrete R-symmetries; no terms of the form $(\overline{27}\ 27)^n$ can appear in the superpotential. So this direction is exactly flat (terms of the form 27^3, $\overline{27}^3$ cannot lift these directions either). In combination with Wilson lines, these flat directions readily break to the $SU(3) \times SU(2) \times U(1)$ group of the Standard Model.

26.6.7 Gauge coupling unification

One of the striking successes of low-energy supersymmetry is its prediction of unification. Within the context of grand unification – where the gauge group of the Standard Model is unified in a simple group at a scale M_{GUT} – the fact that the couplings unify is readily understood. In the context of the compactifications considered here, it is not immediately obvious why this should be the case. In the case of symmetry breaking by Wilson lines, for example, the compactification scale and the scale of the symmetry breaking are of the same order. So there is no energy scale where one has a unified, four-dimensional effective theory.

In the weakly coupled heterotic string, however, the couplings *do unify* under rather broad conditions. In the case of Wilson line breaking this can be understood immediately in field-theoretic terms. The effect of the Wilson line is to eliminate states from the E_6 unified theory, but at tree level no couplings are altered. So the couplings of all groups emerging from E_6 are the same. Perhaps more surprising is the fact that the E_6 and E_8 couplings are the same. This can be seen by considering the vertex operators for the gauge bosons in each group. In both cases, the vertex operators are constructed in terms of free two-dimensional fields, which obey the same algebra (in the unbroken subgroup) as in the flat space theory. So, for example,

the operator product expansions of these gauge boson vertex operators are unaltered. There are constructions where unification does not hold. These involve replacing the two-dimensional fermions with a current algebra with a different central extension.

In the (2, 1) flat directions considered above, we can give an argument based on the low-energy field theory that the couplings remain unified as one moves out along the flat direction. A change in the coupling requires that there be a coupling of this modulus to the gauge fields. But at the classical level, we know that there are no such couplings, because any such coupling would violate the axion shift symmetry. This symmetry is unaffected by the expectation value of these moduli.

When we come to consider strongly coupled strings, the problem of coupling unification will be more complicated. It will be less clear in what sense unification is generic. Whether this is a problem for the theory, or a clue, is a question for the student to ponder.

26.6.8 Calculating the parameters of the low-energy Lagrangian

As we have explained, string theory is a theory without fundamental dimensionless parameters. On the other hand, the structure of the low-energy theory, we now see, depends on discrete choices: which Calabi–Yau, orbifold, etc., in how many dimensions, with how much supersymmetry, with which Wilson lines, and continuous dynamical quantities, the moduli. For any given choice, at least classically, it should be a straightforward problem to calculate the parameters of the low-energy theory.

It is easy to calculate the four-dimensional gauge couplings in terms of the ten-dimensional dilaton and the radius. We have already seen how this works for simple compactifications, and this carries over directly to the Calabi–Yau case, since the vertex operators for the gauge fields are constructed in terms of two-dimensional fields, as in the orbifold or toroidal case.

The cosmological constant is another interesting quantity in the low-energy theory. At the classical level in the Calabi–Yau compactifications, it vanishes. This can be understood in a variety of ways. First, if we examine the solution of the ten-dimensional equations of motion, we see that since the Ricci tensor vanishes, there is no cosmological term. Second, in the two-dimensional conformal field theory, the cosmological constant would give rise to a tadpole for the dilaton, but this is forbidden by conformal invariance. Ultimately, the absence of a cosmological constant is inherent in the form of the solution: we assumed that four dimensions are flat. We will see later that this is not necessary: string theory admits AdS spaces, as well as Minkowski spaces, as classical solutions.

From the perspective of trying to understand the Standard Model, a particularly important set of parameters are the Yukawa couplings. These can certainly be computed in the string theory. In principle, we should construct the vertex operators

for the appropriate matter fields, and then construct the required OPE coefficients or suitable scattering matrices. In practice, this can often be short-circuited. In the orbifold models, for example, in the untwisted sectors we can read off the Yukawa couplings by dimensional reduction of the ten-dimensional Lagrangian. The scalar fields are components of the original ten-dimensional gauge fields, A_i. Similarly the fermions are components of the ten-dimensional gauginos. In the orbifold theory, alternatively, it is not difficult to construct the vertex operators and to compute the required OPE coefficients.

In the Calabi–Yau case, we have seen that, in the α' expansion, the superpotential is independent of R. So one can work at very large radius, and pick off the leading contribution. To actually do the computation, one can construct the zero modes of the scalar and spinor fields, and substitute in the Lagrangian. A priori, one might expect that this would be quite difficult, given that one does not have an explicit formula for the metric. But it turns out that the Yukawa couplings have a topological significance, and their values can be inferred by general reasoning. We will not have use for explicit formulas here, but it is important to be aware of their existence.

26.6.9 Other perturbative heterotic string constructions

The quintic is just one of a large class of Calabi–Yau models which can be constructed. The exact number is not actually known. It is not even known, with certainty, whether the number of Calabi–Yau vacua is finite or infinite.

So while we will not assess here the size of this space, there is clearly a large class of string solutions with gauge group identical to that of the Standard Model. These theories have varying numbers of generations, including both orbifold (or free fermion) models and Calabi–Yau constructions with three. There are many models with groups, numbers of generations, and other features radically different than those of the Standard Model. Still, it is remarkable how easily we have obtained models which accord with some of our speculations for Physics Beyond the Standard Model. We have found low-energy supersymmetry, coupling unification, light Higgs particles, discrete symmetries which can potentially suppress proton decay and give rise to a stable dark matter candidate, all in a framework where we can imagine that real calculations are possible.

In subsequent chapters, we will turn to the problems of actually turning these observations and discoveries into a real theory which we can confront with experiment.

Suggested reading

Volume 2 of Green *et al.* (1987) provides a comprehensive introduction to Calabi–Yau compactification, and I have borrowed heavily from their presentation. Weakly

coupled string models with three generations have been constructed in the context of Calabi–Yau compactification; their phenomenology is considered by Greene *et al.* (1987). Models based on free fermions have been constructed by Faraggi (1999). We will encounter non-perturbative constructions in Chapter 28. At special points in their moduli spaces, some Calabi–Yau spaces can be described in terms of solvable conformal field theories. This program was initiated by Gepner (1987), and is described at some length by Polchinski (1998). A very accessible description, including computations of physically interesting couplings, appears in Distler and Greene (1988).

Exercises

(1) Write the field strength of electrodynamics as a two-form, and express its gauge invariance in the language of forms. Verify that $dF = 0$ is the Bianchi identity (the homogeneous Maxwell equations).

(2) Show that for a Kahler manifold, the non-vanishing components of the affine connection (Christoffel symbols) are given by Eq. (26.36). Then show that the non-zero components of the Riemann tensor are given by Eq. (26.37) and verify Eq. (26.38). Derive Eq. (26.40) by noting that

$$\Gamma^{\bar{a}}_{\bar{b}\bar{a}} = \partial_{\bar{b}} \ln \det g. \tag{26.76}$$

Show that our result for the two-dimensional curvature of a Riemann surface is a special case of this.

(3) For a flat, two-dimensional torus, introduce complex coordinates and verify that the bosonic and fermionic terms are just those of the free string action. You can take $K = X^\dagger X$ for this case.

(4) Write out the action of the heterotic string propagating in the Calabi–Yau background with spin connection equal to the gauge connection in some detail. Determine the form of the vertex operators for the 27 and $\overline{27}$ fields, in the RNS formulation (you can limit yourself to the NS–NS sector).

(5) Exhibit a combination of Wilson lines and $SU(5)$ singlet expectation values which break the gauge group to that of the Standard Model in the case of the quintic in CP_4.

27

Dynamics of string theory at weak coupling

In previous chapters, we have seen that string theory at the classical level shows promise of describing the Standard Model, and can realize at least one scenario for physics beyond: low-energy supersymmetry. But there are many puzzles, most importantly the existence of moduli and the related question of the cosmological constant. At tree level, in the Calabi–Yau solutions, the cosmological constant vanishes. But whether this holds in perturbation theory and beyond requires understanding of the quantum theory.

In studying string theory, we have certain tools:

(1) weak coupling expansions,
(2) long-wavelength (low-momentum, α') expansions.

We have exploited both of these techniques up to this point. In analyzing string spectra, we have worked in a weak coupling limit. There are corrections to the masses and couplings, for example, in string perturbation theory, and most of the states that we have studied have finite lifetimes. At weak coupling, these effects are small, but at strong coupling, the theories presumably look dramatically different.

In asserting that Calabi–Yau vacua are solutions of the string equations, we used both types of expansions. We wrote the string equations both in lowest order in the string coupling, and also with the fewest number of derivatives (two). Even at weak coupling and in the derivative expansion, we can ask whether Calabi–Yau spaces are actually solutions of the string equations, both classically and quantum mechanically. For example, we have seen that, at lowest order in both expansions, there are typically many massless particles. We might expect tadpoles to appear for these fields, both in the α' and in loops. There is in general no guarantee we can find a sensible solution by simply perturbing the original one.

Yet there are many cases where we can make exact statements. In both Type II and heterotic string theories, we can often show that Calabi–Yau vacua correspond

to exact solutions of the classical string equations. We can also show that they are good vacua – there are no tadpoles for massless fields – to all orders of the string perturbation expansion. More dramatically, we can sometimes show that these vacua are good, non-perturbative states of the theory. This is perhaps surprising, since we lack a suitable non-perturbative formulation in which to address this question directly. The key to this magic is supersymmetry. In the framework of quantum field theory, we have already seen that supersymmetry gives a great deal of control over dynamics, both perturbative and non-perturbative. We were able to prove a variety of non-renormalization theorems from very simple starting points. The more supersymmetry, the more we could establish. The same is true in string theory. We can easily prove a variety of non-renormalization theorems for string perturbation theory. We can show that with $N = 1$ supersymmetry in four dimensions, the superpotential is not renormalized from its tree level form in perturbation theory; the gauge coupling functions are not renormalized beyond one loop. These same considerations indicate the sorts of non-perturbative corrections which can (and do) arise. In theories with more supersymmetries, one can prove stronger statements: the superpotential is not renormalized at all, and there are strong constraints on the kinetic terms. These sorts of results will be important when we try to understand weak–strong coupling dualities.

27.1 Non-renormalization theorems

In each of the superstring theories one can prove a variety of non-renormalization theorems. Consider, first, the case of ten dimensions. At the level of two derivative terms, the actions with $N = 1$ or $N = 2$ supersymmetry (16 or 32 supercharges) are unique. So, *perturbatively and non-perturbatively, there is no renormalization.* This is a variant of our discussion in four-dimensional field theories. If we dimensionally reduce the Type II theories on a six-dimensional torus, we obtain a four-dimensional theory with 32 supercharges ($N = 8$ in four dimensions); if we reduce the heterotic theory, we obtain a theory with $N = 4$ supersymmetry in four dimensions (16 supercharges). In either case, the supersymmetry is enough to prevent corrections to either the potential or the kinetic terms, not only perturbatively but non-perturbatively.

These are quite striking results. From this we learn that the question of whether the universe is four-dimensional or not, or whether it has, say, four or eight supersymmetries, or none, is not simply a dynamical question (at least in the naive sense of comparing the energies of different states or their relative stability). Other issues, perhaps cosmological, must come into play. We will save speculations on these questions for later.

27.1.1 Non-renormalization theorems for world sheet perturbation theory

Let us turn, now, to compactified theories. Consider, first, a Type II theory compactified on a Calabi–Yau space. In this case, the low-energy theory has $N = 2$ supersymmetry. Again, this is enough to guarantee that there is no potential generated for the moduli, perturbatively or non-perturbatively. In other words, starting with a solution of the equations of the low-energy effective field theory, at lowest order in g_s and R^2/α', we are guaranteed that we have an exact solution to all orders – and non-perturbatively – in both parameters.

Now consider the compactification of the heterotic string theory on the same Calabi–Yau space, with spin connection equal to the gauge connection. Then the world sheet theory, as we saw, has two left-moving and two right-moving supersymmetries. It is identical to the theory which describes the corresponding Type II background. But we just established that the Calabi–Yau space is a solution of the classical string equations, which means that there is a corresponding superconformal field theory with central charge $c = 9$. This is an exact statement; so the background corresponds to an exact solution of the classical string equations. This does *not* establish that the Calabi–Yau space corresponds to an exact vacuum quantum mechanically, as it does in the Type II case. For example, the intermediate states in quantum loops in the two theories are different.

We can establish this result in a different way. Consider the $h_{1,1}$ $(1, 1)$-forms, $b^{(a)}_{i\bar{i}}$; one of these is the Kahler form, where $b_{i\bar{i}} = g_{i\bar{i}}$. In world sheet perturbation theory we have seen that these fields decouple at zero momentum. The fact that all scattering amplitudes involving external b particles vanish at zero momentum has consequences for the structure of the low-energy effective Lagrangian: only derivatives of b appear in the Lagrangian. This is reminiscent of the couplings of Goldstone bosons; the Lagrangian, in world sheet perturbation theory, is symmetric under

$$b(x) \to b(x) + \alpha \tag{27.1}$$

for constant α.

This result implies a non-renormalization theorem for σ-model perturbation theory; b lies in a supermultiplet with r^2, the modulus which describes the size of the Calabi–Yau space. This is apparent from the fact that they are both Kaluza–Klein modes associated with the metric, $g_{i\bar{i}}$; r^2 is the symmetric part; b is the antisymmetric part. So this is similar to the situation in which we could prove non-renormalization theorems in field theory. Different orders of σ-model perturbation theory are associated with different powers of r^{-2}. But in holomorphic quantities such as the superpotential and gauge coupling function, additional powers of r^{-2} are accompanied by powers of b. So only terms which are independent of r^{-2} are

permitted by the shift symmetry. As a result, the superpotential computed at lowest order is not corrected in σ-model perturbation theory. This means that particles which are moduli at the leading order in α' are moduli to all orders of σ-model perturbation theory.

This non-renormalization theorem does not quite establish that these are good solutions of the classical string theory; there is still the possibility that non-perturbative effects in the σ-model will give rise to potentials for the lowest-order moduli. Indeed, our argument for the vanishing of the b couplings is not complete. At zero momentum, the vertex operator for b, V_b, is topological; while it is the integral of a total divergence, it does not necessarily vanish. There generally exist classical Euclidean solutions of the two-dimensional field theory – instantons – for which the vertex operator is non-zero. These world sheet instantons raise the possibility that non-perturbative effects on the world sheet will lift some or all of the vacuum degeneracy. For the $(2, 2)$ theories, however, we already know that this does not occur. We earlier argued, by considering the compactification of the related Type II theories, that these corresponding sigma models are exactly conformally invariant. It is possible (and not terribly difficult) by examining the structure of the two-dimensional instanton calculation ("world-sheet instanton") to show that no superpotential is generated. While we will not review this analysis here, the techniques involved are familiar from our discussion of four-dimensional instantons. One wants to determine whether instantons can generate a superpotential. One needs, as in four dimensions, to count fermion zero modes, and see if they can lead to a non-vanishing correlation function at zero momentum for an appropriate set of fields. In the $(2, 2)$ case, one finds that they cannot. One can then ask whether quantum corrections (small fluctuations) to the instanton result can yield such a correction. Here, one notes that, as in perturbation theory, holomorphy fixes uniquely the dependence on the coupling. So if the lowest-order contribution vanishes, higher orders vanish as well.

In the case of $(2, 0)$ compactifications of the heterotic string, the situation is more complicated. Perturbatively, we can argue, as before, that solutions of the string equations at lowest order are solutions to all orders in the α' expansion. Non-perturbatively, however, the situation is less clear. For such compactifications, there is no corresponding Type II compactification, so we can not rely on the magic of $N = 2$ supersymmetry. It is necessary to examine in detail the effects of world sheet instantons. In general, if one does the sort of zero-mode counting described above, one finds that it is possible to generate a superpotential. But in many cases, one can argue that there are cancellations, and the superpotential vanishes.

It is important to understand that the non-renormalization theorems do not imply that the Calabi–Yau manifold is itself an exact solution to the classical string equations; rather, the point is that there is guaranteed to exist a solution nearby. There

can be – and are – tadpoles for massive particles in σ-model perturbation theory. A tadpole corresponds to a correction of the equations of motion:

$$\nabla^2 h + m^2 h = \Gamma. \tag{27.2}$$

This is solved by a perturbatively small shift in the h field:

$$h = -\frac{\Gamma}{m^2}. \tag{27.3}$$

For the massless fields, however, one cannot find a solution in this way, and in general, if there is a tadpole, there is no nearby (static) solution of the equations. This is why the low-energy effective action is such a useful tool in addressing such questions: it is precisely the tadpoles for the massless fields which are important.

27.1.2 Non-renormalization theorems for string perturbation theory

In field theory, we proved non-renormalization theorems by treating couplings as background chiral fields, and exploring the consequences of holomorphy of the effective action as functions of these fields. In string theory, we have no coupling constants, but the moduli determine the effective couplings, and since they are themselves fields, they are restricted by the symmetries of the theory. We exploited this connection in the previous section to prove non-renormalization theorems for σ-model perturbation theory. In this section, we prove similar statements for string perturbation theory.

We begin with the heterotic string theory, on a Calabi–Yau manifold or an orbifold. In this case, we saw that there is a field, S, which we called the dilaton (it is sometimes called the four-dimensional dilaton). The vertex operator for the imaginary part of this field, $a(x)$, at $k = 0$, is simply:

$$V_a = \int d^2\sigma \epsilon^{ab} \partial_a X^\mu \partial_b X^\nu b_{\mu\nu}. \tag{27.4}$$

This is, again, a total derivative on the world sheet. So this particle, which we saw earlier is an axion, decouples at zero momentum. So again there is a shift symmetry – this is just the axion shift symmetry. Again, this means that the superpotential must be independent of S. But since powers of perturbation theory come with powers of S, this establishes that the superpotential is not renormalized to all orders of perturbation theory!

As in the world sheet case, there can be non-perturbative corrections to the superpotential, and this raises the possibility that potentials will be generated for the moduli. We will see shortly that gluino condensation, as in supersymmetric field theories, is one such effect.

First, we consider other string theories. In the case of Type II compactified on a Calabi–Yau space, the $N = 2$ supersymmetry is enough to insure that no superpotential is generated perturbatively or non-perturbatively: Calabi–Yau spaces correspond to exact ground states of the theory, and the degeneracies are exact as well. As in field theories with $N = 2$ supersymmetry, corrections to the metric (Kahler potential) are possible. Theories with more supersymmetry (heterotic on tori or Type II theories on $K3$ spaces with $N = 4$ supersymmetry, or Type II on tori with eight supersymmetries) are even more restricted.

27.2 Fayet–Iliopoulos D-terms

In deriving the non-renormalization theorems for string perturbation theory, we established that there is no renormalization of the superpotential, or of the gauge coupling function beyond one loop. But this is not quite enough to establish that there is no renormalization of the *potential*. We must also check whether Fayet–Iliopoulos terms are generated. From field-theoretic reasoning, we might guess that any renormalization would occur only at one loop. In globally supersymmetric theories in superspace, a Fayet–Iliopoulos term has the form:

$$\zeta^2 D = \int d^4\theta V. \tag{27.5}$$

This term is just barely gauge invariant; under $V \to V + \Lambda + \Lambda^\dagger$, this is invariant because $\int d^4\theta \Lambda = 0$ since Λ is chiral. If we treat the gauge coupling (or any other couplings) as background fields, any would-be corrections to D would have the form:

$$\int d^4\theta g(S, S^\dagger) V \tag{27.6}$$

which is only invariant if g is a constant. Thus any D-term is independent of the coupling, in the normalization where $1/g^2$ appears in front of the gauge terms. So at most there is a one-loop correction.

Before going on to string theory, it is interesting to look at the structure of any one-loop term. Call the associated $U(1)$ generator Y. If supersymmetry is unbroken, massive fields come in pairs with opposite values of Y, so only massless fields contribute. The Feynman diagram which contributes to the D-term is shown in Fig. 27.1. It is given by:

$$\zeta^2 = \mathrm{Tr}(Y) \int \frac{d^4k}{(2\pi)^4} \frac{1}{k^2}. \tag{27.7}$$

So a vanishing D-term requires that the trace of the $U(1)$ generator vanish. The one-loop diagram is quadratically divergent, but let's rewrite this in a way which

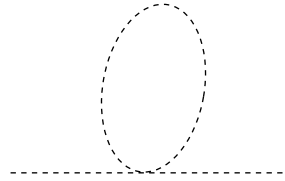

Fig. 27.1. The Feynmann diagram which contributes to the D-term.

resembles expressions we have seen in string theory. We can introduce a "Schwinger parameter," which we will call τ_2. Then:

$$\begin{aligned}\zeta^2 &= 2\pi \mathrm{Tr}(Y) \int_0^\infty d\tau_1 \int \frac{d^4k}{(2\pi)^4} e^{-2\pi\tau_2 k^2} \\ &= \frac{1}{32\pi^3}\mathrm{Tr}(Y) \int_0^\infty \frac{d\tau_2}{\tau_2^2} \int_{-1/2}^{1/2} d\tau_1 .\end{aligned} \tag{27.8}$$

We have written things in this way because we want to think think of this as an integral over the modular parameter of the torus. At this stage, the integral is still quadratically divergent. But, under modular transformations, the complex τ plane is mapped into itself several times. We can define a *fundamental domain*,

$$-\frac{1}{2} \le \tau_1 \le \frac{1}{2}, \quad |\tau| \ge 1. \tag{27.9}$$

If we restrict the integration to the fundamental domain, the result is finite. In string theories, this turns out to be the correct answer:

$$\zeta^2 = \frac{1}{192\pi^2}\mathrm{Tr}\ Y. \tag{27.10}$$

This result can be derived by a straightforward string computation. But instead, in string models where Tr(Y) is non-zero, we can give a low-energy field theory argument which completely fixes the coefficient of the D-term, and also sheds light on possible perturbative corrections. If $\mathrm{Tr(Y)} \neq 0$, the low-energy theory has a gravitational anomaly. This anomaly is rather similar to the gauge anomalies we have discussed in field theory. It arises from a diagram with one external gauge boson and an external graviton. String models with such anomalies typically have gauge anomalies as well, which we can readily evaluate. As an example, consider the compactification of the $O(32)$ heterotic string on a Calabi–Yau space, with spin connection equal to the gauge connection. In this case, the low-energy gauge group is $SO(26) \times U(1)$. There are $h_{1,1}$ 26s with $U(1)$ charge 1, and $h_{2,1}$ 26s with $U(1)$ charge -1. There are also corresponding singlets, with charge $+2$ and -2 respectively. These are in precise correspondence with the fields we found in E_6; the 26s arise in parallel to the $O(10)$ 10s; the singlets to the $O(10)$ singlets. But

now it is clear that there are anomalies in the gauge symmetries. For example, there is a $U(1) \times O(26)^2$ anomaly proportional to

$$A = (h_{2,1} - h_{1,1}) \tag{27.11}$$

and a $U(1)^3$ anomaly:

$$A' = (h_{2,1} - h_{1,1})(26 - 8). \tag{27.12}$$

On the other hand, this is a modular invariant configuration of string theory, so there should not be any inconsistency, at least in perturbation theory. So something must cancel the anomaly. The cancellation is actually a variant of the mechanism discussed originally by Green and Schwarz in ten dimensions, now specialized to four dimensions. We know that there is a coupling:

$$\int d^2\theta S W_\alpha^2. \tag{27.13}$$

This gives rise to a coupling of the axion to the $F\tilde{F}$ terms of each group. The anomaly calculation in the low-energy theory implies a variation of the action proportional to the anomaly coefficient and $F\tilde{F}$. So if the axion transforms under the gauge symmetry as

$$a(x) \to a(x) + c\omega(x) \tag{27.14}$$

this can cancel the anomaly. It is crucial that the anomaly coefficients are the same for each group.

We can check whether this hypothesis is correct. If $a(x)$ transforms, then it must couple to the gauge field. The required covariant derivative is

$$D_\mu a = \partial_\mu a - \frac{1}{c} A_\mu. \tag{27.15}$$

So from the kinetic term in the action, there is a coupling of A_μ to a. One can compute this coupling without great difficulty and verify that it has the required magnitude.

More interesting, however, is to consider the implications of supersymmetry. We can generalize the coupling above to superspace. The transformation law for a now becomes a transformation law for S:

$$S \to S + \Lambda + \Lambda^\dagger \tag{27.16}$$

where Λ is the chiral gauge transformation parameter. The gauge-invariant action for S is:

$$-\int d^4\theta \ln\left(S + S^\dagger - \frac{1}{c} V\right). \tag{27.17}$$

If we Taylor series expand this Lagrangian, we see that, in addition to the $A_\mu \partial^\mu a$ coupling, we generate a Fayet–Iliopoulos D-term:

$$\int d^4\theta \frac{1}{c(S+S^\dagger)} V. \tag{27.18}$$

One can verify that this term – and the other terms implied by this analysis – are present. First, we can ask: at what order in perturbation theory should each of these terms appear? To establish this, we need to remember that the standard supergravity Lagrangian is written in a frame where $M_{\rm p}^2$ appears in front of the Einstein term in the effective action. In the string frame, it is the dilaton – essentially S – which appears out front. If we rescale the four-dimensional metric by:

$$g_{\mu\nu} \to S g_{\mu\nu} \tag{27.19}$$

then S appears in front of the Lagrangian. With this same rescaling, the "kinetic" term, which had an S out front, has S^3. The Fayet–Iliopoulos D-term, originally has a $1/S$ out front. Correspondingly, the resulting scalar mass term would be proportional to $1/S^2$. After the metric rescaling, this is independent of S, i.e. in the heterotic string theory, the D-term should appear at one loop, in accord with our field theory intuition. Similarly, the coupling $A_\mu \partial^\mu a$ should appear at one loop, while there should be a contribution to the cosmological constant at two loops. All of these can be found by straightforward string computations (some of these are described in the Suggested reading).

In essentially all known examples, this one-loop D-term does not lead to supersymmetry breaking. There always seem to be fields which can cancel the D-term. Consider, again, the $O(32)$ theory. Here we can try to cancel the D-term by giving an expectation value to one of the singlets, 1_{-2}. The question is whether this gives a non-zero contribution to the potential when we consider the superpotential. The most dangerous coupling is a term $1_{-2}1_{+2}$ involving some other singlet. But such terms are absent at lowest order, and their absence to higher orders is guaranteed by the non-renormalization theorems. Charge conservation forbids terms of the form 1_{-2}^n; there are no other dangerous terms. So this corresponds to an exact "F-flat" direction of the theory. So in perturbation theory there exists a good vacuum. While a general argument is not known, empirically this possibility for cancellation appears to arise in every known example.

What does the theory look like in this new vacuum?

(1) Supersymmetry is restored and the vacuum energy vanishes.
(2) The $U(1)$ gauge boson has a mass-squared of order g_s^2 times the string scale.
(3) The longitudinal mode of the gauge boson is principally the imaginary part of the charged scalar field whose vev canceled the D-term. There is still a light axion.

From the perspective of a very-low-energy observer, the D-term is not a dramatic development. It plays some role in determining physics at a very-high-energy scale (albeit not quite as high as the string scale). What is perhaps most impressive is the utility of effective field theory arguments in sorting out a microscopic string problem. Prior to the discovery of the D-term, for example, there had been many papers "proving" a strict non-renormalization theorem for the *potential*; this, we see, is not correct (it is not hard to determine, in retrospect, what went wrong in the original proofs). The effective field theory arguments make clear when the potential is renormalized in perturbation theory and when it is not. They also permit easily finding the "new vacuum" in cases where a Fayet–Iliopoulos term appears. It is possible, in principle, to find this vacuum by stringy methods, but this is distinctly more difficult. Finally, these arguments give insight into the non-perturbative fate of the non-renormalization theorems.

27.3 Gaugino condensation

We have seen that in string theory, if supersymmetry is unbroken at tree level, it is unbroken to all orders of perturbation theory. The argument, as in field theory, allows exponential dependence on the coupling. In the case of the heterotic string compactified on a Calabi–Yau space, gaugino condensation, as in supersymmetric field theories, generates a superpotential on the moduli space.

Consider the $E_8 \times E_8$ theory compactified on a Calabi–Yau space, with spin connection equal to the gauge potential, and without Wilson lines. In this case, there is an $E_6 \times E_8$ gauge symmetry. There are typically several fields in the 27 of E_6, but there are no chiral fields transforming in the E_8. So one has a pure E_8 supersymmetric gauge theory. The couplings of the E_6 and E_8 are equal at the high scale, so the E_8 coupling becomes strong first. This leads, as we have seen, to gaugino condensation. We have also seen that at tree level there is a coupling:

$$S W_\alpha^2. \tag{27.20}$$

Just as before, this leads to a superpotential for S:

$$W(S) = A e^{-3S/b_0}. \tag{27.21}$$

One often hears this described as a "field theory analysis," as if it is not necessarily a feature of the string theory. But string theory obeys all of the principles of quantum field theory. If we correctly integrate out high-energy string effects, the *low-energy analysis is necessarily reliable.* So the only question is: are there terms in the low-energy effective action that lead to larger effects. One might worry that, since we understand so little about non-perturbative string theory, it would be hard to address

this. But with some very mild assumptions, we can establish that the *low-energy effects are parametrically larger than any high-energy string effects.*

The basic assumption is that, as in field theory, non-perturbatively the theory obeys a discrete shift symmetry (for a suitable normalization of a):

$$a(x) \to a(x) + 2\pi. \tag{27.22}$$

When we discuss non-perturbative string theory, we will give some evidence for this; this will turn out to be one of the more mild assertions in the subject of string duality. For now, note that if we accept this, any superpotential for S arising from high-energy string effects is of the form

$$W_{\rm np} = C_n e^{-nS} \tag{27.23}$$

for integer n. So such effects are exponentially smaller than gaugino condensation.

What does the low-energy theory look like? The dilaton potential goes rapidly to zero for large S, i.e. in the weak coupling limit. We might have hoped that somehow we would find supersymmetry is broken and the moduli are fixed. But instead, gaugino condensation leads to a runaway potential. At large S, we have just argued that no additional stringy effects can stabilize this behavior.

We can imagine more elaborate versions of this phenomenon, involving matter fields as well, in some sort of hidden sector. But it is difficult to construct models where the moduli are stabilized in any controlled fashion along these lines.

27.4 Obstacles to a weakly coupled string phenomenology

We have seen that string theory is a theory without dimensionless parameters. This is an exciting prospect, but it also raises the question: how are the parameters of low-energy physics determined? We have argued that the answer to this question lies in the dynamics of the moduli: the expectation values of these fields determine the couplings in the low-energy Lagrangian.

In non-supersymmetric string configurations, perturbative effects already lift the degeneracy among different vacua, giving rise to a potential for the moduli. In the previous section, we have learned that in supersymmetric compactifications non-perturbative effects generically lift the flat directions of the potential. In other words, the moduli are not truly moduli at the quantum level. At best, we can speak of approximate moduli in regions of the field space where the couplings are weak. The potentials, both perturbative and non-perturbative, all tend to zero at zero coupling. This is not surprising; with a little thought, it becomes clear that this behavior is not specific to perturbation theory or some particular non-perturbative phenomenon such as gaugino condensation. At very weak coupling, we expect that the potential always tends rapidly to zero. This means that if the potential has a minimum, this

occurs when the coupling is not small. This is troubling, for it means that it is likely to be hard – if possible at all – to do computations which will reveal detailed features of the state of string theory (if any) which describes the world we see around us.

In the next chapter, we will see that much is known about non-perturbative string physics. Most striking are a set of dualities, which relate regimes of very strong coupling in one string theory to weak coupling in another. While impressive, these, by themselves, do not help with the strong coupling problem we have elucidated above. If, at very strong coupling, the theory is equivalent to a weakly coupled theory, the potential will again tend to zero. In other words, it is likely that stable ground states of string theory exist only in regions where no approximation scheme is available.

Perhaps just as troubling is the problem of the cosmological constant. Neither perturbative nor non-perturbative string theory seems to have much to say. The potentials are more or less of the size one would guess from dimensional analysis (and the expected dependence on the coupling). Perhaps most importantly, they are, up to powers of coupling, as large as the scale set by supersymmetry breaking.

There are, however, some reasons for optimism. Perhaps the most important is provided by nature itself: the gauge and Yukawa couplings of the Standard Model are small. Another is provided by string theory. As we will discuss later, there are ways in which large pure numbers can arise. These might provide mechanisms to understand the smallness of couplings, even in situations where asymptotically the potential vanishes. Finally, we will see that there is, at present, only one proposal to understand the smallness of the cosmological constant, and string theory may provide a realization of this suggestion.

Suggested reading

The result that there are no continuous global symmetries in string theory is a fundamental one. For the heterotic theory, it appears in Banks and Dixon (1988). Non-renormalization theorems for world sheet perturbation theory and issues in construction of (0, 2) models are described by Witten (1986) and by Green *et al.* (1987). The non-renormalization theorem for string perturbation theory is described by Dine and Seiberg (1986). The space-time argument for the Fayet–Iliopoulos D-term appears in Dine *et al.* (1987c); world sheet computations appear in Atick *et al.* (1987) and Dine *et al.* (1987a). World sheet instantons are discussed in Dine *et al.* (1986, 1987b); cancellations of instanton effects relevant to (0, 2) theories are studied by Silverstein and Witten (1995).

28

Beyond weak coupling: non-perturbative string theory

In the previous chapter, we were forced to face the fact that string theory, if it describes nature, is not weakly coupled. On the other hand, the very formulation we have put forward of the theory is perturbative. We have described the quantum mechanics of single strings, and given a prescription for calculating their interactions order by order in perturbation theory in a parameter g_s. There is a parallel here to Feynman's early work on relativistic quantum theory: Feynman guessed a set of rules for computing perturbative amplitudes of electrons. In this case, however, one already had a candidate for an underlying description: quantum electrodynamics. It was Dyson who clarified the connection. For Abelian theories, the non-perturbative theory probably does not really exist, but in the case of non-Abelian gauge theories it does. The field theoretic formulation provides an understanding of the underlying symmetry principles, and access to a trove of theoretical information.

A string field theory would be a complicated object. The string fields themselves would be functionals of the classical two-dimensional fields which describe the string. The quantization of such fields is sometimes called "third quantization." Much effort has been devoted to writing down such a field theory. For open strings, one can write relatively manageable expressions which reproduce string perturbation theory. For closed strings, infinite sets of contact interactions are required. But apart from their cumbersome structure, there are reasons to suspect that this is not a useful formulation. There would seem to be, for example, vastly too many degrees of freedom. At one loop, we have seen that string amplitudes are to be integrated only over the fundamental region of the moduli space. Naively a field theory which simply describes all of the states of the string would have amplitudes integrated over the whole region, and the cosmological constant would be extremely divergent. The contact terms mentioned above solve this problem, but not in a very satisfying way.

Despite this, there has been great progress in understanding non-perturbative aspects of the known string theories. Most strikingly, it is now known that all theories

with 16 or more supersymmetries are the same. Many tools have been developed to study phenomena beyond string perturbation theory, especially *D*-branes and supersymmetry. There exist some cases where non-perturbative formulations of string theory are possible, and we will discuss them briefly in this chapter. They are technically and conceptually much simpler than string field theory. They have a puzzling, perhaps disturbing feature, however: they are special to strings propagating in particular backgrounds. It is as if, in Einstein's theory, for each possible geometry, one had to give a different Hamiltonian. All of these results are "empirical." They have been developed by collecting circumstantial evidence on a case-by-case basis. There is still much which is not understood. In the last chapter, we will discuss how this developing understanding might lead to a closer connection of string theory to nature.

28.1 Perturbative dualities

Before considering examples of weak–strong coupling dualities, we return to the large/small radius duality we studied in Section 25.3. Many of the dualities we will study have a similar flavor, even though they cannot be demonstrated so directly. We saw that there is an equivalence of the heterotic string theory at small radius to the theory at large radius. By examining the action of these transformations at their fixed points, we saw that these duality symmetries are gauge symmetries. We could ask, as well, the significance of duality transformations in the IIA and IIB theories. As in other closed strings, in addition to transforming the radii, the duality transformation takes:

$$\partial X^9 \to -\partial X^9; \quad \bar{\partial} X^9 \to \bar{\partial} X^9. \tag{28.1}$$

Because of world sheet supersymmetry, it has the same action on the fermions; $\psi^9 \to -\psi^9$; $\tilde{\psi}^9 \to \tilde{\psi}^9$. But under this the chirality operator appearing in the GSO projector is reversed in sign, i.e. duality interchanges the IIA and IIB theories; *the small-radius IIA theory is equivalent to the large-radius IIB theory*, and vice versa. There are other perturbative connections. For example, the compactified $O(32)$ heterotic string theory is equivalent to the $E_8 \times E_8$ theory.

28.2 Strings at strong coupling: duality

Duality is a term used in physics to label different descriptions of the same physical situation. At the level of perturbation theory, we have learned about five apparently different string theories. Based on the perturbative dualities discussed above, we see that there are at most three inequivalent string theories, the Type I, Type II, and heterotic theories. But it is tempting to ask whether there are more connections. In

this chapter, we will see that all of the known string theories are equivalent in a similar way, but these equivalences relate small and large coupling. For example, the strong coupling limit of the $O(32)$ heterotic string theory is the weak coupling limit of Type I string theory; the strongly coupled limit of $E_8 \times E_8$, compactified to six dimensions on a torus, is the weakly coupled limit of Type II theory compactified on a K3 manifold (K3 manifolds are essentially four-dimensional Calabi–Yau spaces); the ten-dimensional Type II theory is self-dual, and, perhaps most intriguingly of all, the strong coupling limit of Type IIA theory in ten dimensions is described, at low energies, by a theory whose low-energy limit is eleven-dimensional supergravity.

Lacking a non-perturbative formulation of the theory, the evidence for these connections is necessarily circumstantial. While circumstantial, however, it is compelling. All of the evidence relies on supersymmetry. We will not be able to review all of this here, but will try to give the flavor of some of the arguments. Supersymmetry, especially supersymmetry with 16 or 32 supercharges, allows one to write a variety of exact formulas, for Lagrangians (based on strong non-renormalization theorems) and for spectra (based on BPS formulas) which can be trusted in both weak and strong coupling limits. This allows detailed tests of the various dualities.

28.3 *D*-branes

When we discussed strong–weak (electric–magnetic) dualities in field theory, topological objects played a crucial role. The same is true in string theory, where the solitons are various types of branes. In general, a *p-brane* is a soliton with a $p+1$-dimensional world volume, so a 0-brane is a particle, a 1-brane a string, a 2-brane a membrane, and so on. In general, one might construct these by solving complicated non-linear differential equations. But a large and important class of topological objects can be uncovered in string theory in a different – and much simpler – way. These are the D-branes. These branes fill an important gap in our understanding of the Type I and Type II theories. In these theories, we encountered gauge fields in the Ramond–Ramond sectors: two-forms in Type I, one-forms and three-forms in the IIA theories, zero-forms, two-forms, and four-forms in the IIB. One natural question is: where are the charged objects that couple to these fields? They are not within the perturbative string spectrum. The vertex operators for these fields involved the gauge-invariant field strengths only, so in perturbation theory there are no objects with minimal coupling. The answer is the D-branes. Their masses (tensions) are proportional to $1/g_s$, so at weak coupling they are very heavy. This is why they are not encountered in the string perturbation expansion.

When we discussed open strings, we noted that there are two possible choices of boundary conditions: Neumann and Dirichlet. At first sight, Neumann boundary

conditions appear more sensible; Dirichlet boundary conditions would violate translational invariance, implying that strings end at a particular point(s). But we have already encountered violations of translational invariance within translationally invariant theories: solitons, such as magnetic monopoles or higher-dimensional objects like cosmic strings and domain walls. Admitting the possibility of Dirichlet boundary conditions for some or all of the coordinates leads to a class of topological objects known as D-branes (for Dirichlet branes). If $d - p - 1$ of the boundary conditions are Dirichlet, while $p + 1$ are Neumann, the system is said to describe a Dp-brane.

We can be quite explicit. Start first with the bosonic string. For the Neumann directions, we have our previous open string mode expansion of Eq. (21.16). For the Dirichlet directions, we have:

$$X^I = x_0^I + i \sum_{n \neq 0} \frac{1}{n} \alpha_n^I e^{-in\tau} \sin(n\sigma) \quad I = 1, d - p - 1. \tag{28.2}$$

Note that there are no momenta associated with the Dirichlet directions. The x_0^Is should be thought of as collective coordinates. We will argue shortly that the tension of the branes is proportional to M_s^{p+1}/g_s.

Consider an "extreme" case, that of a $D0$-brane. There are 25 collective coordinates and no momenta, so this object is a conventional soliton. In field theory, the excitations near the soliton, which describe scattering of mesons (field theory excitations) from the soliton must be found by studying the eigenfunctions of the quadratic fluctuation operator. But here they are very simple: they are just the excitations of the open string. As a second example, consider a $D3$-brane. Now the momentum has four components. So the excitations which propagate on the brane are four-dimensional fields. These break up into two types. The Neumann fields, X^μ, give rise to a massless gauge boson, the state $\alpha_{-1}^\mu|0\rangle$; the Dirichlet fields, X^I, give rise to massless scalars on the brane $\alpha_{-1}^I|0\rangle$. In the superstring version of this construction, there are six scalars and a gauge boson, and their superpartners. In $N = 1$ language, this is a vector multiplet and three chiral multiplets, the content of $N = 4$ Yang–Mills theory with gauge group $U(1)$.

Before considering some of these statements in greater detail, let us explore a further aspect of this construction. Suppose we have several branes, say $D3$-branes, parallel to each other. Here parallel just means that the strings which end on these branes have Dirichlet or Neumann boundary conditions in the same direction. Now, however, we have the possibility that the strings end on different branes. Take the simplest case of two branes. If the branes are separated by a distance r, in addition to the modes above, labeled by the collective coordinate x_i^I, $i = 1, 2$, we have to

allow for expansions of the form:

$$X^I(\sigma,\tau) = x_i^I + \sigma\frac{r}{\pi}\left(x_j^I - x_i^I\right) + i\sum_{n\neq 0}\frac{1}{n}\alpha_n^I e^{-in\tau}\sin(n\sigma) \quad I = 1,\ldots,d-p-1. \tag{28.3}$$

There are two such configurations, one starting on the first brane and ending on the second, and one starting on the second and ending on the first. The ground states in these sectors have mass-squared proportional to r^2. For $r \neq 0$, all of these states are massive. The massless bosons consist of a $U(1)$ gauge boson on each brane, as well as scalars. As $r \to 0$, we have two additional massless gauge bosons. If we generalize to n branes, we have n massless gauge bosons and $6n$ scalars; as we bring the branes close together, we have n^2 gauge bosons and $6n^2$ scalars.

There is a natural conjecture as to what is going on here. When all of the branes are on top of one another, we have a $U(n)$ gauge symmetry, with $3n$ complex scalars transforming in the adjoint representation of the group. As the branes are separated, the adjoint scalars acquire (commuting) expectation values; these break the gauge symmetry to $U(1)^n$, giving mass to the other gauge bosons. In principle, we would like to check that these n^2 gauge bosons interact as required for Yang–Mills theories, as we did for the gauge bosons of the heterotic string. This is more challenging here, since we need vertex operators which connect strings ending on different branes and we will not attempt this. We will provide further evidence for the correctness of this picture shortly.

The branes break some of the supersymmetry of the Type II theory in infinite space; instead of 32 conserved supercharges, there are 16. A simple way to understand this uses the light cone gauge construction. There are now open strings ending on the brane. For the world sheet fermions, the boundary conditions relate the left and right movers on the string. Calling these S_a and $\tilde{S}_a$, we have

$$S^a(\sigma,\tau) = \sum_n S_n^a e^{-in(\tau+\sigma)} \quad \tilde{S}^a(\sigma,\tau) = \sum_n S_n^a e^{-in(\tau-\sigma)}. \tag{28.4}$$

Recall that half of the supercharges have the very simple form:

$$Q^a = \int d\sigma\, S^a \quad \tilde{Q}^a = \int d\sigma\, \tilde{S}^a \tag{28.5}$$

so $Q^a = \tilde{Q}^a$. This is the structure of a broken supersymmetry generator, with S the goldstino. The same is true for the other set of supercharges. Other configurations of branes, such as non-parallel sets of branes, preserve less supersymmetry. Brane–anti-brane configurations preserve no supersymmetry at all.

We can imagine other sets of branes, which would respect different amounts of supersymmetry. If we have branes which are not parallel, for example, different sets

of supersymmetries will be preserved. In order to count supersymmetries, we need to compare the supersymmetries on different branes, at different angles relative to one another.

28.3.1 Brane charges

We have seen that the simplest D-brane configurations preserve half of the supersymmetries. In other words, they are BPS states. Typically BPS states are associated with conserved charges. In the case of IIA and the IIB theories, in the Ramond–Ramond sectors there are gauge fields, but, in perturbation theory, no charged objects. Polchinski guessed – and showed – that the objects which carry Ramond–Ramond charges are D-branes. In the IIA case, the gauge fields are a one form and a three form; in the IIB case they are a zero-form, a two-form, and a (self-dual) four-form. In relativistic mechanics, a gauge field couples to a particle – a zero-brane. We have seen that a two-index tensor couples naturally to a string – a one-brane. So this suggests that in the IIA theory, there should be Dp-branes with p even, coupling to the corresponding R–R gauge fields, while in the IIB theory there should be Dp-branes with p odd. Polchinski verified this by direct calculation. He computed the one-loop amplitude for two separated branes. For large separations, he found the poles associated with exchange of the massless gauge fields (more precisely, for fixed separation, r, one should see falloff with powers of $1/r$). His calculation not only yields the brane charges, but it also gives the brane tensions.

Consider the case of two branes, separated by a distance y. In empty flat space, the trace over states in the one-loop amplitude for open strings gives a result of the form:

$$\mathcal{A} = C \int_0^\infty \frac{dt}{t^2}. \tag{28.6}$$

The power of t arises from the momentum integral, $\int d^8k \, \exp(-k^2)$, as well as from the manipulation of the oscillator traces. The main difference in the case of two separated branes is that the mass-squared has a contribution from the brane separation, y^2, and $9-p$ coordinates of the brane are fixed, so they don't have associated momenta. So the result has the form:

$$\begin{aligned} \mathcal{A} &= C \int_0^\infty \frac{dt}{t^2} (8\pi^2 \alpha' t)^{(9-p)/2} \exp\left(-\frac{t y^2}{2\pi\alpha'}\right) \\ &\sim y^{-(7-p)} \sim G_{9-p}(y). \end{aligned} \tag{28.7}$$

Here $G_d(y)$ is the scalar Green function in d dimensions. So one can think of a potential between the branes associated with the exchange of massless states. These massless states are antisymmetric tensor fields and their superpartners, as well as gravitons and gravitinos. These contributions can be isolated, and the tensions and

charges of the D-branes determined. In the case of the superstring, the full potential vanishes due to boson and fermion cancellations.

28.3.2 Brane actions

We are familiar with the actions for zero-branes and one-branes. The action for a general p-brane is a generalization of these:

$$S_p = -T_p \xi \int d^{p+1}\xi \det\left(\frac{\partial X^\mu}{\partial \xi^a}\frac{\partial X^\nu}{\partial \xi^b}\eta_{\mu\nu}\right)^{1/2}. \tag{28.8}$$

In the zero-brane case, this is the action for a particle; $X^\mu(\tau)$ is the collective coordinate which describes the position of the soliton, and T_0 is its mass. For a general background with a bulk metric, dilaton and antisymmetric tensor field this generalizes to

$$S_p = -T_p \int d^{p+1}\xi e^{-\Phi}[-\det(G_{ab} + B_{ab} + 2\pi\alpha' F_{ab})]^{1/2}. \tag{28.9}$$

The terms involving the metric and antisymmetric tensor are similar to those we have encountered elsewhere in string theory, and their form is not surprising. The $e^{-\Phi}$ reflects the fact that in the open string sector, the coupling constant is the square root of that of the closed string sector.

28.4 Branes from T-duality of Type I strings

There is another way to think about D-branes, which provides further insight. We have seen that closed string theories exhibit a duality between large and small radius. In the heterotic theory there is an exact equivalence of the theories at large and small radius, which can be understood as a gauge symmetry. In Type II theories, T-duality relates two apparently different theories. It is natural to ask what is the connection between large and small radius in theories with open strings. Open strings have momentum states, but no winding states. So there cannot be a self-duality. Instead, we look for an equivalence between the open string theory at one radius and some other theory at the inverse radius. Here we uncover D-branes.

Consider the boundary conditions on the strings in the compactified direction. For the closed string fields, the effect of the duality transformation is to take:

$$X_{\rm L} \to X_{\rm L} \quad X_{\rm R} \to -X_{\rm R}. \tag{28.10}$$

In terms of left- and right-moving bosons in open string theories, Neumann boundary conditions are the conditions

$$\partial_\tau X = (\partial_{\sigma_+} + \partial_{\sigma_-})X = 0. \tag{28.11}$$

So after a T-duality transformation, we would expect

$$(\partial_{\sigma_+} - \partial_{\sigma_-})X = \partial_\sigma X = 0, \tag{28.12}$$

i.e. we have traded Neumann for Dirichlet boundary conditions. While this follows from simple calculus manipulations, it is instructive to formulate this in terms of the mode expansion for the open string. Ordinarily, we have:

$$X^9 = x_i^9 + \frac{1}{2}p(\tau+\sigma) + \frac{1}{2}p(\tau-\sigma) + i\sum_{n\neq 0}\frac{1}{n}\left(\alpha_n^9 e^{-in(\tau+\sigma)} + \alpha_n^9 e^{-in(\tau-\sigma)}\right). \tag{28.13}$$

The effect of the duality transformation is to change the sign of the terms which depend on $\tau - \sigma$. So instead of writing an expansion in terms of cosines, we have an expansion in terms of sines:

$$X^9 = x_0^9 + p\sigma + i\sum_{n\neq 0}\frac{1}{n}\alpha_n^9 e^{-in\tau}\sin(n\sigma). \tag{28.14}$$

These are precisely the Dirichlet branes. Note the role of p: in the T-dual picture, it is a sort of winding: it describes strings which start on the brane, wind around the compact dimension some number of times, and then end on the brane.

This T-duality of open strings also allows us to better understand the appearance of gauge interactions associated with stacks of branes. In the original open string picture, gauge degrees of freedom are described by Chan–Paton factors, i.e. charges on the ends of the string. In the case of Type I strings, these are described by states of the form $|AB\rangle$, $A, B = 1, \ldots, 32$. Now consider a $U(16)$ subgroup of $O(32)$. The string ends carry labels, i, j, within $U(16)$. Taking the diagonal generators of $U(N)$ to be the matrices

$$T_1 = \mathrm{diag}(1, 0, 0, \ldots) \quad T_2 = \mathrm{diag}(0, 1, 0, \ldots), \tag{28.15}$$

etc., the state $(\bar{i}, j)$ carries charge -1 under T_i, $+1$ under T_j, and zero under the other generators.

We can consider constant, background gauge fields in the 9 direction. We can write these as:

$$A = \mathrm{diag}(a_1, a_2, \ldots, a_{16}). \tag{28.16}$$

This has a gauge-invariant description in terms of the Wilson line:

$$U = e^{i\oint d\vec{x}\cdot\vec{A}}, \tag{28.17}$$

where the integral is taken in the periodic directions. Such a background gauge field breaks the gauge symmetry to $U(1)^{16}$, in general; the other gauge bosons should

gain mass. In field theory, the corresponding mass terms are proportional to

$$[A^\mu, A^9]^2 \tag{28.18}$$

so the diagonal gauge bosons are massless, and those corresponding to the non-Hermitian generator

$$T_{ij}^{kl} = \delta_i^k \delta_j^l \tag{28.19}$$

have mass-squared

$$m^2 = (a_i - a_j)^2. \tag{28.20}$$

This is similar to the calculations we did of symmetry breaking in grand unified theories.

We would like to understand how this result arises directly in the string theory. It is simplest to consider the case of a string which is constant in σ. The coupling of the string depends on the Chan–Paton factors. In the light cone, the action in the presence of a gauge field is like that of a particle:

$$\frac{1}{2}\int d\tau \left(\left(\frac{\partial \dot{X}^9}{\partial \tau} \right)^2 + (a_i - a_j) \frac{\partial X^9}{\partial \tau} \right). \tag{28.21}$$

For a non-constant string, the situation is somewhat more complicated, since the gauge fields couple at the string end points.

The extra term modifies the canonical momenta. These are now:

$$P = \frac{n}{R} = \frac{\partial \dot{X}^9}{\partial \tau} + (a_i - a_j). \tag{28.22}$$

This means that the leading term in the string mode expansion is:

$$X^9 = \left(\frac{n}{R} - (a_i - a_j) \right) \tau. \tag{28.23}$$

This gives an extra contribution to the mass. If $n = 0$, this is *exactly* what we expect from field-theoretic reasoning.

Now let's consider the T-dual picture. Under T-duality, the zero-mode part of X transforms into:

$$X^9 = x_0 + \left(\frac{n}{R} - (a_i - a_j) \right) \sigma. \tag{28.24}$$

For $i = j$, this corresponds to a string beginning and ending on the same D-brane. For $i \neq j$, the string ends at different points, i.e. on separated D-branes. At least for the Type I theory, we have *derived* the picture we conjectured earlier: a stack of N coincident branes describes a $U(N)$ gauge symmetry; as the branes are separated, the gauge symmetry is broken by a field in the adjoint representation.

28.4.1 Orientifolds

We have seen that we can understand the appearance of D-branes by considering T-duality transformations of open strings. The Type I theory is a theory of *oriented* strings. In the closed string sector, the action has a parity symmetry, which interchanges left and right on the world sheet. Calling the corresponding operator Ω, one keeps only states which are invariant under the action of Ω. This is necessary for the consistency of interactions of open and closed strings. This means that closed string states like

$$\alpha_{-2}\tilde{\alpha}_{-1}\tilde{\alpha}_{-1}|0\rangle \tag{28.25}$$

are not allowed, but symmetrized combinations such as

$$\alpha_{-2}\tilde{\alpha}_{-1}\tilde{\alpha}_{-1} + \tilde{\alpha}_{-2}\alpha_{-1}\alpha_{-1}|0\rangle \tag{28.26}$$

are permitted. This projection is similar to the orbifold projections we have encountered earlier.

Consider the action of Ω in the T-dual theory. We have seen that in terms of the original fields,

$$X^{9\prime} = -X^9_{\mathrm{L}} + X^9_{\mathrm{R}}. \tag{28.27}$$

So the effect of interchanging left and right is to change the sign of X^9, i.e. Ω is a combination of a world sheet parity transformation and a reflection in space-time.

The effect of this projection on states is similar to a Z_2 orbifold projection. We can combine momentum states to form states with definite transformation properties under the reflection:

$$|p\rangle \pm |-p\rangle. \tag{28.28}$$

Gravitons, for example, in the non-compact directions, $G_{\mu\nu}$, must have momentum states which are even; in coordinate space, this means that graviton states must be even functions of x. The fields $G_{\mu 9}$ must be odd functions, and so on. It is as if there is an entity, the *orientifold*, sitting at the origin – the fixed point of the reflection. This object in fact has a negative tension. One way to see this is simply to note that the effect of the T-duality transformation was to produce a set of D-branes. These branes have a positive tension. From the point of view of the non-compact dimensions, this is a cosmological constant. But the original theory had no such cosmological constant – this must be canceled by the orientifold.

Just as it is not necessary to start from the Type I theory and its dualities to encounter D-branes, it is not necessary to start from the Type I theory to consider orientifolds. Starting from Type II theories, in particular, we can perform a projection by world sheet parity times some Z_2 space-time symmetry. For example,

consider a Type II theory with a single compact dimension. On this theory, we can make a projection which is a combination of world sheet parity, Ω, and reflection in the compact dimension.

28.5 Strong–weak coupling dualities: the equivalence of different string theories

We have seen that at weak coupling, there are a variety of connections between different string theories which are surprising from a field-theoretic perspective. The heterotic string, compactified on a circle of very large radius, is equivalent to a string theory compactified at very small radius (with a different coupling). The Type IIA theory at large radius is equivalent to the IIB theory at small radius. The $O(32)$ heterotic string is equivalent to the $E_8 \times E_8$ theory. All of these equivalences involve significant rearrangement of the degrees of freedom. Typically Kaluza–Klein modes, which are readily understood from a space-time field theory point of view, must be exchanged with winding modes, which seem inherently "stringy." So perhaps it is not surprising that there are other equivalences, involving weak and strong coupling. Again, we have had some inkling of this in field theory, when we studied $N = 4$ Yang–Mills theory. There, the theory at weak coupling is equivalent to a theory at strong coupling. To see this equivalence, one needs to significantly rearrange the degrees of freedom. States with different electric and magnetic charge exchange roles, as the coupling is changed from strong to weak.

In string theory, there is a complex web of dualities. The IIB theory in ten dimensions exhibits a strong–weak coupling duality very similar to that of $N = 4$ Yang–Mills theories; weak and strong coupling are completely equivalent. The $O(32)$ heterotic string theory, in ten dimensions, is equivalent at strong coupling to the weakly coupled Type I theory. These relations are surprising, in that these theories appear to involve totally different degrees of freedom at weak coupling. But there are more surprises still. The strong coupling limit of the IIA theory in ten dimensions is a theory whose low-energy limit is eleven-dimensional supergravity. If we allow for compactifications of the theory, this set of dualities is already enough to establish an equivalence of all string theories, as well as some as yet not fully understood theory whose low-energy limit is eleven-dimensional supergravity. But as we compactify, we find further, intricate relations. For example, the Type IIA theory on $K3$ is equivalent to $E_8 \times E_8$ on T^4. Given that all of the sensible theories of quantum gravity we know are equivalent, it is plausible that, in some sense, there is a unique theory of quantum gravity. As we will see, however, we only know this reliably for theories with at least 16 supercharges. For theories with four or less, this situation is less clear; it is by no means obvious that the statement is even meaningful.

In the next sections, we will explore some of these dualities, and the evidence for them. We will also discuss two particularly surprising equivalences. We will argue that certain string theories are equivalent to *quantum field theories* – even to quantum mechanical systems. The very notion of space-time in this framework will be a derived concept.

28.6 Strong–weak coupling dualities: some evidence

In the case of T-dualities, those dualities which relate the behavior of string theories at weak coupling and different radii, it is straightforward to understand the precise mappings between the different descriptions. Lacking a general non-perturbative definition of string theory, it is not possible to do something similar in the case of strong–weak coupling dualities. Instead, one can try to put together compelling circumstantial evidence. Without supersymmetry, even this is essentially impossible. But in the presence of sufficient supersymmetry one has a high degree of control over the dynamics. Evidence for equivalence can be provided by studying the following.

(1) The effective action: in ten or eleven dimensions, the terms in the action with up to two derivatives are uniquely determined by supersymmetry, so they are not corrected either perturbatively or non-perturbatively. A similar statement holds for $N \geq 4$ actions in four (and actions with varying degrees of supersymmetry in between). In some cases, one can check higher-derivative terms in the effective action as well.
(2) The spectrum of BPS objects: in many cases, the low-lying states are BPS objects. They cannot disappear from the spectrum as the coupling or other parameters are varied. With 16 or more supercharges, they obey *exact* mass formulae. The identity of the BPS states for different theories provides non-trivial evidence for these equivalences.

We will explore only some of the simplest connections here, but it is important to stress that these identifications are often subtle and intricate. In many instances where one might have thought the dualities mentioned above might fail, they do not.

28.6.1 IIA → eleven-dimensional supergravity (M theory)

We start with the IIA theory, where we can readily access both aspects of the duality. Comparing the actions of eleven-dimensional supergravity and the IIA theory is particularly straightforward, as the Lagrangian of the IIA theory is often obtained by compactifying eleven-dimensional supergravity on a circle, keeping only the zero modes. The basic degrees of freedom in eleven dimensions are the graviton, g_{MN}, the antisymmetric tensor gauge field, C_{MNO}, and the gravitino, ψ_M.

We are not going to work out the detailed properties of this theory, but it is a useful exercise to check that the numbers of bosonic and fermionic degrees of freedom are the same. As usual, we can count degrees of freedom by going to the light cone (or using the "little group," the group of rotations in $D = 11 - 2 = 9$). The metric is a symmetric, traceless tensor; for the gravitino, we need also to impose the constraint $\gamma^I \psi_i = 0$. For the metric, then, we have $((9 \times 10)/2) - 1 = 44$, while from the three-index antisymmetric tensor we have $(9 \times 8 \times 7)/3! = 84$, for a total of 128 bosonic degrees of freedom. From the gravitino, we have $9 \times 16 - 16 = 128$ degrees of freedom.

If we compactify x^{10} on a circle of radius R, we obtain the following bosonic degrees of freedom in ten dimensions.

(1) The ten-dimensional metric, $g_{\mu\nu}$ ($\mu, \nu = 0, \ldots, 9$).
(2) From $g_{10\mu}$ we obtain a vector gauge field. This is identified with the Ramond–Ramond vector field of the IIA theory.
(3) From $C_{10\mu\nu}$ we obtain an antisymmetric tensor field, identified with the antisymmetric tensor, $B_{\mu\nu}$ of the NS–NS sector of the IIA theory.
(4) From $C_{\mu\nu\rho}$, we obtain the three-index antisymmetric tensor field of the R–R sector of the IIA theory.
(5) From $g_{10,10}$ we obtain a scalar field in ten dimensions, the dilaton of the IIA theory. Note that this mode corresponds to the radius, R, of the eleventh dimension.

Now consider the action. We will examine just the bosonic terms. These are constructed in terms of the curvature tensor, the three-index antisymmetric tensor, and its corresponding four-index field strength, F:

$$\mathcal{L} = \frac{-1}{2\kappa^2}\sqrt{g}\mathcal{R} - \frac{1}{48}\sqrt{g}F^2_{MNPQ} - \frac{\sqrt{2}\kappa}{3456}\epsilon^{M_1\ldots M_{11}}F_{M_1\ldots M_4}F_{M_5\ldots M_8}C_{M_9M_{10}M_{11}}. \tag{28.29}$$

As we indicated, the dimensional reduction of this theory gives the Lagrangian of the IIA theory in ten dimensions. It is convenient to parameterize the fields in terms of the vielbein, e^A_M. Then:

$$e^A_M = \begin{pmatrix} e^A_\mu & A_\mu \\ 0 & R_{11} \end{pmatrix}. \tag{28.30}$$

Correspondingly, the metric has the structure:

$$g_{MN} = e^A_M e^B_N \eta_{AB} = \begin{pmatrix} g_{\mu\nu} & R_{11}A_\mu \\ R_{11}A_\nu & R^2_{11} \end{pmatrix}. \tag{28.31}$$

If we simply plug these expressions into the Lagrangian, the coefficient of the Einstein, $\mathcal{R}$, term, will be proportional to R. In order to bring this Lagrangian to the

canonical, Einstein form, it is necessary to perform a Weyl rescaling of the metric. But instead, we will perform the rescaling so as to bring the action to "string frame." In this frame, all of the NS–NS fields have a factor of $e^{-2\phi}$ out front, where $e^{-2\phi}$ is the string coupling (dilaton). In ten dimensions, $\sqrt{g} = e$ transforms like $(g_{\mu\nu})^5$ under an overall rescaling of the metric; $\mathcal{R}$ transforms like $(g_{\mu\nu})^{-1}$. So we need to rescale:

$$g_{\mu\nu} \to R_{11}^{-2/3} g_{\mu\nu}. \tag{28.32}$$

The three form, C, upon reduction, leads to various fields in ten dimensions. The components $C_{10\mu\nu}$ give the NS–NS two-form. The fields $C_{\mu\nu\rho}$ give the R–R three-form. The R–R one-form field arises from the $g_{10,\mu}$ components of the metric. The ten-dimensional action becomes:

$$S = S_{\rm NS} + S_{\rm R} \tag{28.33}$$

with

$$S_{\rm NS} = \frac{1}{2}\int d^{10}x\ \sqrt{g} e^{-2\phi}\left(\mathcal{R} + (\nabla\phi)^2 - \frac{1}{2}H^2\right) \tag{28.34}$$

$$I_{\rm R} = -\int d^{10}x\ \sqrt{g}\left(\frac{1}{4}F^2 + \frac{1}{2\times 4!}F_4^2\right) - \frac{1}{4}F_4 \wedge F_4 \wedge B. \tag{28.35}$$

We have seen that, when the action is written this way, R is related to the coupling of the ten-dimensional string theory. The Weyl rescaling, $g_{\mu\nu} \to R_{11}^{-3/4} g_{\mu\nu}$ gives an action with R^3 out front, i.e.

$$\mathcal{L} = R_{11}^{-3}\left(-\frac{1}{2}\mathcal{R} - \frac{3}{4}R^{-3/2}H_{\mu\nu\rho}^2 - \frac{9}{16}\left(\frac{\partial_\mu R_{11}}{R_{11}}\right)^2\right). \tag{28.36}$$

In this form, the unit of length is the string scale, $\ell_{\rm s}$. So loops come with a factor of R_{11}^3 (the ultraviolet cutoff is $\ell_{\rm s}^{-1}$). So we see that

$$g_{\rm s}^2 = \frac{R_{11}^3}{\ell_{11}^3}. \tag{28.37}$$

We can derive this relation another way (not keeping $2\pi s$), which makes a more direct connection between eleven-dimensional supergravity and strings. The eleven-dimensional theory has membrane solutions. We will not exhibit these here, but this should not be too surprising: the three-form, C_{MNO}, couples naturally to membranes. The eleven-dimensional theory has only one scale, ℓ_{11}, so the tension of the membranes is of order ℓ_{11}^{-3}. We can wrap one of the coordinates of the membrane around the eleventh dimension. If the eleventh dimension is very small,

the result is a string propagating in ten dimensions, with a tension:

$$T = \ell_{11}^{-3} R = \ell_{\rm s}^{-2}. \tag{28.38}$$

Now, again, the ten-dimensional gravitational coupling is related to ℓ_{11} by

$$G_{10} = \frac{\ell_{11}^9}{R_{11}}. \tag{28.39}$$

So we find, again,

$$g_{\rm s}^2 = \frac{R_{11}^3}{\ell_{11}^3}. \tag{28.40}$$

So we have our first piece of circumstantial evidence for the connection. Let's turn now to the BPS spectrum. Consider, first, the eleven-dimensional supersymmetry algebra. Eleven-dimensional spinors can be decomposed into ten-dimensional spinors of definite chirality, with indices α and $\dot{\alpha}$. In this basis,

$$\Gamma_{11} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{28.41}$$

The eleven-dimensional momenta decompose into ten-dimensional momenta and p_{11} in an obvious way:

$$\{Q_\alpha, Q_{\dot{\alpha}}\} = \not{p}_{\alpha,\dot{\alpha}} + p_{11}\delta_{\alpha,\dot{\alpha}}. \tag{28.42}$$

From a ten-dimensional point of view, the last term is a central charge. In the presence of such a central charge, we can prove a BPS bound as we did for the monopole. This bound is saturated by the Kaluza–Klein modes of the graviton and the antisymmetric tensor field. What charge does this central charge correspond to in the IIA theory, and to which states do the momentum states correspond? It is natural to guess that this is one of the R–R charges. The simplest possibility is the charge associated with the one-form gauge field. The carriers of the one-form charge are $D0$-branes. The $D0$-branes are BPS states – they preserve half of the ten-dimensional supersymmetry. So *states of definite eleven-dimensional momentum are states of definite D-brane charge.* More precisely, localized states with N units of Kaluza–Klein momentum correspond to zero-energy bound states (so-called threshold bound states) of N D-branes.

There are numerous further tests of this duality. For example, if one compactifies the theory further, there are connections to IIB theory. There are also connections involving $M5$-branes. But this discussion gives some flavor of the duality, and the evidence.

28.6.2 IIB self-duality

The IIB theory exhibits an interesting self-duality. We can understand this, first, from the Lagrangian. The Lagrangian for the NS–NS fields is the same as for the IIA theory. For the R–R fields, we have now zero and two, and four-form fields. The Lagrangian for these is similar, with appropriate indices, to that for the R–R fields of the IIA case. A careful examination shows that under the transformation $\phi \to -\phi$ the Lagrangian goes into itself. At the classical level, the action is also invariant under shifts of the axion.

Grouping the dilaton, e^{ϕ}, and the Ramond–Ramond scalar, θ, into a complex field,

$$\tau = \frac{4\pi i}{g_s} + \frac{\theta}{2\pi}, \tag{28.43}$$

it then is natural to conjecture that the underlying theory has an $SL(2, Z)$ symmetry similar to that of $N = 4$ Yang–Mills theory:

$$\tau \to \frac{a\tau + b}{c\tau + d} \quad ad - bc = 1. \tag{28.44}$$

Further evidence for this symmetry is obtained by studying BPS objects: the various branes of the theory. In the IIB theory, we have fundamental strings and $D1$-branes; we also have $D5$-branes. Under this duality, the fundamental strings are mapped into $D1$-branes by the $SL(2, Z)$ transformations. Correspondingly, the H_3 form (which couples to fundamental strings) should be mapped into the F_3 form (which couples to $D1$ strings). The $D3$-branes are associated with the gauge-invariant five-form field strength, which is self-dual, so we might expect the $D3$-branes to be invariant. Study of the BPS formulae for these states lends support to these conjectures.

This leaves the $D5$-branes. These couple to the Ramond–Ramond six-form form gauge field, which is associated with a seven-form field strength, which is in turn dual to the three form R–R field strength. In other words, the $D5$-brane is a magnetic source for F_3. So we might expect these to be dual to something which is a magnetic source for the NS three form. This would be an NS 5-brane. Such an object can be constructed as a soliton of the ten-dimensional IIB supergravity theory. It plays an important role in understanding the duality of these theories. It also appears in other contexts. For example, in M theory, it is associated with a seven-form field strength, which is dual to the four-form field strength we have already encountered.

The $M5$ solution is:

$$g_{mn} = e^{2\phi}\delta_{mn};\ g_{\mu\nu} = \eta_{\mu\nu} \tag{28.45}$$

$$H_{mno} = -\epsilon_{mno}{}^{p}\partial_p\phi \tag{28.46}$$

$$e^{2\phi} = e^{2\phi(\infty)} + \frac{Q}{2\pi^2 r^2}. \tag{28.47}$$

Here μ, ν are the coordinates tangent to the brane (they are the world-volume coordinates); $m, n \ldots$ are the coordinates transverse to the brane. The $SL(2, Z)$ duality of the IIB theory is quite intricate and beautiful. There are many subtle and interesting checks.

28.6.3 Type I–$O(32)$ duality

The duality between the Type I and $O(32)$ theories is particularly intriguing, as it is a duality between a theory with open and closed strings and a theory with closed strings only. It is also puzzling since the perturbative spectra of these theories, at the level of massive states, are quite different. The $O(32)$ heterotic theory contains towers of massive states in spinor representations; there is nothing like this in the perturbative spectrum of the Type I theory. By way of evidence, we can begin, again, with the effective Lagrangian. For the heterotic theory this can be written

$$\int d^{10}x e^{-2\phi}(\mathcal{R} + |\nabla\phi|^2 + F^2 + dB^2). \tag{28.48}$$

Here $e^{-2\phi}$ is the dilaton field, and we have written the action in string frame. Consider, now, the transformation:

$$g = e^{\phi}g' \quad \phi = -\phi'. \tag{28.49}$$

This takes the action to:

$$\int d^{10}x\sqrt{g}(e^{-2\phi'}(R + |\nabla\phi'|^2) + e^{-\phi'}F^2 + dB^2). \tag{28.50}$$

This is the action for the bosonic fields of the Type I theory. The closed string fields couple with g^2, while the open string fields couple with g. In the Type I theory, the antisymmetric tensor is an R–R field, and as a result, no factor of the coupling (the dilaton) appears out front of its kinetic term.

Now we can ask: how do the hetorotic strings appear in the open string theory? Here, we might guess that these strings would appear as solitons. More precisely, these strings are just the $D1$-branes of the Type I theory. At weak coupling, the tension of these strings will behave as $1/g$, i.e. it will be quite large. In this sector, one can find states in spinorial representations of $O(32)$, arising from configurations

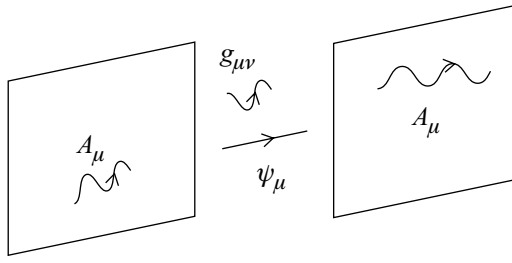

Fig. 28.1. The strongly coupled heterotic string is described by an eleven-dimensional bulk theory and two segregated walls, on which gauge degrees of freedom propagate.

of $D1$–$D9$-branes. Most important, the $D1$-branes are BPS. As a result, they persist to strong coupling, and in this regime their tension is small. We will not explore the various subtle tests of this correspondence, but other features one can investigate include the identification of the winding strings of the heterotic theory.

Many other dualities among different string theories have been explored. These include an equivalence between heterotic string theory on a four-torus and Type IIA on K3, and equivalences of Calabi–Yau compactifications of the Type II theory and heterotic theory on $K3 \times T2$.

28.7 Strongly coupled heterotic string

In ten dimensions, we have seen that the strong coupling limit of the IIA theory is a theory whose low-energy limit is eleven-dimensional supergravity. The strong coupling limit of the IIB theory is again the IIB theory. The strong coupling limit of the $O(32)$ heterotic string is the Type I string. This still leaves the question: what is the strong coupling limit of the $E_8 \times E_8$ heterotic string? The answer is intriguing. It has some tantalizing connections to facts we see in nature. It also suggests different ways of thinking about compactifications – inklings of the large extra dimension and warped space pictures which we will discuss in the next chapter.

Horava and Witten recognized that the strong coupling limit of the heterotic string, like the IIA theory, is an eleven-dimensional theory. The theory is defined on an interval of radius R_{11}. The relation of R_{11} to the string tension and coupling are exactly as in the IIA case. This means that as the coupling gets large, the interval becomes large. We will refer to the full eleven-dimensional space as the "bulk." The fields propagating in the bulk are a full eleven-dimensional supergravity multiplet: graviton, gravitino and three-form field. At the end of the interval, there are two walls (Fig. 28.1). These walls are similar to orientifolds, in that they are not dynamical (there are no degrees of freedom corresponding to motion of the walls). The

low-lying degrees of freedom on each wall are those of a supersymmetric E_8 gauge theory: gauge bosons and gauginos in the adjoint representation. The Lagrangian has the structure of a bulk plus boundary term:

$$S = -\frac{1}{2\kappa^2}\int d^{11}x\sqrt{g}\mathcal{R} - \sum_{i=1}^{2}\frac{1}{8\pi}(4\pi\kappa^2)^{2/3}\int d^{10}x\sqrt{g}\mathrm{Tr}F_i^2 + \cdots . \quad (28.51)$$

Note that the gauge coupling is simply proportional to the sixth power of the eleven-dimensional Planck length.

Support for this picture comes from a variety of sources. First, there is a subtle cancellation of gauge and gravitational anomalies. Second, the long-wavelength limit of this theory is ten-dimensional gravity plus Yang–Mills theory, with the relation between the gauge and gravitational couplings appropriate to the heterotic string (this is one way to fix the coupling constants). Further compactifications provide further checks.

28.7.1 Compactification of the strongly coupled heterotic string

One puzzle in the phenomenology of the weakly coupled heterotic string concerns the value of the gauge coupling and the unification scale. In the MSSM, the unification scale is two orders of magnitude below the Planck scale. If we imagine that the unification scale corresponds to a scale of compactification, then

$$\alpha_{\mathrm{gut}} \propto \frac{g_{\mathrm{s}}^2}{V}. \quad (28.52)$$

If we treat the left hand side as fixed, then as V becomes large, so does g_{s}. Plugging in the observed values, g_{s} is quite large. As we will now show, the situation in the strong coupling limit is much different – and much more promising.

Consider compactification of the strongly coupled theory on a Calabi–Yau space. The full compact manifold, from the point of view of an eleven-dimensional observer, is the product of the interval times a Calabi–Yau space X. Such a configuration is an approximate solution of the lowest-order equations of motion. Even at the level of the classical equations of this theory, there are corrections arising from the coupling of bulk to boundary fields. These can be constructed in a power series expansion. Terms in the expansion grow with R_{11}, owing to the one-dimensional geometry in the eleventh dimension. They are proportional to $\kappa^{2/3}$, from the bulk–brane coupling in Eq. (28.51). On dimensional grounds, there is a factor of R^{-4}, where r is the Calabi–Yau radius. The expansion parameter is thus

$$\epsilon = \kappa^{2/3}R_{11}/R^4. \quad (28.53)$$

We can readily obtain the relation between the four-dimensional and eleven-dimensional quantities. Using the string relations (here we will be careful about factors of 2 and π):

$$G_N = \frac{e^{2\phi}(\alpha')^4}{64\pi V} \quad \alpha_{\text{gut}} = \frac{e^{2\phi}(\alpha')^3}{16\pi V} \tag{28.54}$$

where V is the volume of the compact space X, and the eleven-dimensional relations:

$$G_N = \frac{\kappa^2}{16\pi^2 V R_{11}} \quad \alpha_{\text{gut}} = \frac{(4\pi\kappa^2)^{2/3}}{2V}, \tag{28.55}$$

we have:

$$R_{11}^2 = \frac{\alpha_{\text{gut}}^3 V}{512\pi^4 G_N^2} \quad M_{11} = R^{-1}\left(2(4\pi)^{-2/3}\alpha_{\text{gut}}\right)^{-1/6}. \tag{28.56}$$

where $R = V^{1/6}$. Putting in the "observed" value of α_{gut} and the four-dimensional Planck mass gives:

$$R_{11}M_{11} = 18 \quad R = 2\ell_{11} = (3 \times 10^{16})\ \text{GeV}. \tag{28.57}$$

The regime of validity of the strongly coupled description is the regime where V and R_{11} are large compared to ℓ_{11}. We see that nature might well be in such a regime. If we evaluate the expansion parameter ϵ, we find $\epsilon \sim 1$. Adopting the viewpoint that the ground state of string theory which describes nature should be strongly coupled, this, again, seems promising: the parameters of grand unification correspond to the point where the eleven-dimensional expansion is just breaking down, $\epsilon \approx 1$. This is in contrast to the weak coupling picture, which seems far from its range of validity.

Apart from this rather direct phenomenological application of string theory ideas, there are two new possibilities which this analysis suggests. First, some compact dimensions might be large compared to the Planck scale (or any fundamental scale). Second, in a case with a one-dimensional geometry, this dimension can be significantly warped, i.e. the metric need not be a constant. These ideas underlie the large extra dimension and Randall–Sundrum models of compactification, which we will encounter in the next chapter.

28.8 Non-perturbative formulations of string theory

We have seen that, at least in cases with a great deal of supersymmetry, we have a surprisingly large access to non-perturbative dynamics. But much of the evidence for the various phenomena we have described is circumstantial, matching actions and spectra in various regions of a given string moduli space. We lack a general,

non-perturbative formulation of the theory, analogous to, say, the lattice formulations of Yang–Mills theories which we encountered in Part 1. One might have hoped that there would be a *string field theory*, analogous to ordinary quantum field theories, but such a program is fraught with conceptual and technical difficulties. We have mentioned some of these. In this section, we will describe situations where one can give a complete non-perturbative description. These descriptions are specific to particular backgrounds: flat space in higher dimensions, and certain AdS spaces. In eleven dimensions, the flat space, supersymmetric theory can be described as an ordinary quantum mechanical system, while the theory compactified on an n-dimensional torus is described by a field theory in $n+1$ space-time dimensions, up to $n=3$. Quite generally, string theory (gravity) in AdS spaces is described by *conformal* field theories; this is known as the AdS–CFT correspondence. Both formulations exhibit what is believed to be a fundamental feature of any quantum theory of gravity: holography. The holographic principle asserts that the number of degrees of freedom of a quantum theory of gravity grows, not as the volume of the system, but as its area.

28.8.1 Matrix theory

We have seen that the strong coupling limit of the IIA theory is an eleven-dimensional theory, whose low-energy limit is eleven-dimensional supergravity. $D0$-branes were crucial in making the correspondence. The Kaluza–Klein states of the eleven-dimensional theory were bound states of $D0$-branes; states with momentum N/R_{11} corresponded to zero-energy ("threshold") bound states of N $D0$-branes. The world-line theory of N $D0$-branes is ten-dimensional $U(N)$ Yang–Mills theory reduced to zero dimensions. The action which describes this system is:

$$S = \int dt \left[\frac{1}{g} \mathrm{tr}(D_t X^i D_t X^i) + \frac{1}{2g} M^6 R_{11}^2 \mathrm{tr}([X^i, X^j][X^i, X^j]) + \frac{1}{g} \mathrm{tr}(i\theta^T D_t \theta + M^3 R_{11} \theta^T \gamma^i [X^i, \theta]) \right], \tag{28.58}$$

where R_{11} is the eleven-dimensional radius, M is the eleven-dimensional Planck mass and $g = 2R_{11}$. The Xs are the bosonic variables, X_I, $I = 1, \ldots, 9$; θs are the fermionic coordinates. It is necessary to impose Gauss's law as a constraint on states.

Classically and quantum mechanically, this system has a large moduli space, corresponding to configurations with commuting X^Is. For large X^I, the spectrum in these directions consists, in the language of quantum mechanics, of $9N$ free particles, and a set of oscillators with frequencies of order $|\vec{X}|$. We can integrate out the fast degrees of freedom, obtaining an effective action for the low-energy degrees

of freedom, the X^Is and their superpartners. The bosonic states are just momentum states for these particles. They are the states corresponding to the collective modes of the D-branes.

Banks, Fischler, Shenker and Susskind made the bold hypothesis of identifying these degrees of freedom, and the Lagrangian of Eq. (28.58), as the complete description of the eleven-dimensional theory, in the limit that $N \to \infty$. More precisely, the Hamiltonian following from the action of (28.58) is identified with the light cone Hamiltonian, and N is identified with the light cone momentum, $P^+ = N/R$. In the large-N limit, this becomes a continuous variable; it is necessary to take $R \to \infty$ at a suitable rate. The first step in this identification is to note that the spectrum of low-lying states of the matrix model is precisely that of the light cone supergravity theory. We have already noted that the states are labeled by a momentum nine-vector, $\vec{p}$. In addition, there are sixteen fermionic variables, the partners of the bosons. As in other contexts, we can define eight fermionic creation operators and eight fermionic destruction operators. From these we can construct a Fock space with 256 states, of which half are space-time bosons (integer spin), and half are fermions. This is just the correct number to describe a graviton and antisymmetric tensor in eleven dimensions, and their superpartners. The states transform correctly under the little group.

A more convincing piece of evidence comes from studying the S-matrix of the matrix theory. Consider, for example, graviton–graviton scattering. Integrating out the massive states of the theory gives an action involving derivatives of x. We won't reproduce the detailed calculation here, but the basic behavior is easy to understand. One can compute the action from Feynman graphs, just as in field theory. With four external Xs, simple power counting gives an action, in coordinate space, behaving as:

$$\mathcal{L}_I \approx \dot{X}^4 \int \frac{dk}{(k^2 + M^2)^4} \cong c\frac{v^4}{M^7}. \tag{28.59}$$

Here $M \propto |X| = R$, the separation of the gravitons. The four factors of v correspond to the four derivatives in the graviton–graviton amplitude; $1/R^7$ is precisely the form of the graviton propagator in coordinate space. With a bit more work, one can show that one obtains precisely the four-graviton amplitude in eleven dimensions, for suitable kinematics.

M theory compactified on an n-torus is described by an $n + 1$-dimensional field theory. We won't argue this, but note that in this case the power counting is correct to give the right graviton–graviton scattering amplitude. If $n > 3$, however, the theory is non-renormalizable, and the description does not make sense. An alternative description can be formulated for dimensions down to six. The matrix model has been subjected to a variety of other tests. It turns out that the large-N limit is not

necessary; for fixed N, one describes a discretized version of the light cone theory (DLCQ). One can actually derive this result, starting with the assumed duality between IIA theory and eleven-dimensional supergravity.

All of this is quite remarkable. Without even postulating the existence of ordinary space-time, we have uncovered space-time, and general relativity, in a simple quantum mechanics model. One interesting feature of these constructions is the crucial role played by supersymmetry. Without it, quantum effects would lift the flat directions and one would not have space-time – though one would still have a sensible quantum system. One might speculate that what we think of as space-time is not fundamental, but almost an accident, associated with the dynamics of particular systems. Lacking, however, a formulation for a realistic, non-supersymmetric system, this remains speculation.

28.8.2 The AdS/CFT correspondence

An equally remarkable equivalence arises in the case of string theory on AdS spaces. This connection was first conjectured by Maldacena, and is referred to as the AdS/CFT correspondence. It asserts that graviton theories in AdS spaces have a description in terms of conformal field theories on the boundary.

A little more general relativity: anti-de Sitter space

We could construct anti-de Sitter space by solving the Freedman equation with negative cosmological constant. Instead, we will adopt a more geometrical viewpoint. Starting with a flat $p+3$-dimensional space, with metric:

$$ds^2 = -dx_0^2 - dx_{p+2}^2 + \sum_{i=1}^{p+1} dx_i^2 \tag{28.60}$$

we consider the hyperboloid:

$$x_0^2 + x_{p+2}^2 - \sum_{i=1}^{p+1} x_i^2 = R^2. \tag{28.61}$$

These coordinates can be parameterized in various ways. For example, one can take

$$\begin{aligned} x_0 &= R\cosh(\rho)\cos(\tau),\ x_{p+2} = R\cosh(\rho)\sin(\tau) \\ x_i &= R\sinh(\rho)\Omega_i\big(i = 1, \ldots p+1;\ \Omega_i^2 = 1\big). \end{aligned} \tag{28.62}$$

This automatically satisfies (28.61), and yields the metric:

$$ds^2 = R^2(-\cosh^2\rho d\tau^2 + d\rho^2 + \sinh^2\rho d\Omega^2). \tag{28.63}$$

In making the correspondence, another parameterization is helpful. These cover one half of the hyperboloid

$$\begin{aligned} x_0 &= \frac{1}{2u}(1 + u^2(R^2 + \vec{x}^2 - t^2)); \; x_{p+2} = Rut \\ x^i &= Rux^i \; (i = 1, \ldots p) \\ x^{p+1} &= \frac{1}{2u}(1 - u^2(R^2 - \vec{x}^2 + t^2)). \end{aligned} \tag{28.64}$$

The metric is then:

$$ds^2 = R^2 \left(\frac{du^2}{u^2} + u^2(-dt^2 + d\vec{x}^2) \right). \tag{28.65}$$

AdS has interesting features, which we will not fully explore here. There is a boundary at spatial infinity ($u = \infty$). Light can reach the boundary in finite time, but not massive particles. In a cosmological context, negative cosmological constant leads not to an eternal AdS space but to a singularity. The last form of the metric will be useful in making the correspondence in a moment. The metric has isometries (symmetries); the group of isometries can be seen from the form of the hyperboloid and the underlying metric of the $p + 3$-dimensional space; it is $SO(2, p + 1)$. This turns out to be the same symmetry as conformal symmetry in $p + 1$ dimensions; this, again, is a crucial aspect of the AdS/CFT correspondence.

Maldacena's conjecture

Maldacena originally discovered this connection for the case of string theory on $AdS_5 \times S_5$. One suggestive argument starts by considering a set of N parallel $D3$-branes. We have discussed such configurations as open string configurations, but they can also be uncovered as solitonic solutions of the supergravity equations, here of the IIB theory. For these, the metric has the form:

$$\begin{aligned} ds^2 &= H(y)^{-1/2} dx^\mu dx_\mu + H(y)^{1/2} \left(dy^2 + y^2 d\Omega_5^2 \right) \\ F_{\mu\nu\rho\sigma\tau} &= \epsilon_{\mu\nu\rho\sigma\tau\alpha} \partial^\alpha H. \end{aligned} \tag{28.66}$$

Here the x^μs are the coordinates tangent to the branes, while the ys (and their associated angles) are the transverse coordinates. The dilaton in this configuration is a constant; the other antisymmetric tensors vanish. The function H, for N parallel branes, is

$$H(\vec{y}) = 1 + \sum_{i=1}^{N} \frac{4\pi g_s (\alpha')^2}{|\vec{y} - \vec{y}_i|^4}. \tag{28.67}$$

This can be rewritten as:

$$ds^2 = \left(1 + \frac{L^2}{y^4}\right)^{-1/2} \eta_{\mu\nu} dx^\mu dx^\nu + \left(1 + \frac{L^2}{y^4}\right)^{1/2} \left(dy^2 + y^2 d\Omega_5^2\right). \quad (28.68)$$

The parameter L is related to the string coupling, g_s, the brane charge (number of branes) N, and the string tension by:

$$L^4 = 4\pi g_s N (\alpha')^2. \quad (28.69)$$

It is convenient to introduce a coordinate $u = L^2/y$, and to take a limit where N and g_s are fixed, while $\alpha' \to 0$. The metric then becomes:

$$ds^2 = L^2 \left[\frac{1}{u^2} \eta_{\mu\nu} dx^\mu dx^\nu + \frac{du^2}{u^2} + d\Omega_5^2\right]. \quad (28.70)$$

The terms involving u and x we have seen previously; this is the geometry of AdS_5. The remaining terms describe a five-sphere of radius L.

Now from a string point of view, the low-energy limit of the system of N $D3$-branes is described by $N = 4$ Yang–Mills theory. So we might, with Maldacena, conjecture that there is just such an equivalence. Not surprisingly, demonstrating this equivalence is not so simple. One needs to argue that on the string side, the bulk modes (graviton, antisymmetric tensors, and so on) decouple, as do the massive excitations of the open strings ending on the branes. One cannot argue this at weak coupling, and it would be surprising if one could; in that case, one could calculate any quantity in the gravity theory in a weak coupling perturbation expansion in the Yang–Mills theory. This is similar to the situation in the matrix model. There are, however (as in the matrix model), many quantities which are protected by supersymmetry, and quite detailed tests are possible, both in this case and for many other examples of the correspondence.

Suggested reading

Non-perturbative string dualities are discussed extensively in the second volume of Polchinski's (1998) book. This provides an excellent introduction to D-branes. D-branes are treated at length in the text by Johnson (2003), as well. The reader may want to consult earlier papers on duality, especially Witten (1995). Matrix theory and the AdS/CFT correspondence are treated in several excellent pedagogical reviews (Bigatti and Susskind, 1997; Aharony *et al.*, 2000; D'Hoker and Freedman, 2002), but the original papers are very instructive; see, for example, Banks *et al.* (1997); Seiberg (1997), Maldacena (1997), Witten (1998).

Exercises

(1) D-branes: for a stack of N D-branes, write the open string mode expansions. Show that, for small separations, the spectrum looks like that of a Higgsed $U(N)$ field theory, with Higgs in the adjoint representation. In the light cone gauge, check the counting of supersymmetries for open strings and D-branes.

(2) Verify the construction of the bosonic terms in the ten-dimensional action from the dimensional reduction of the eleven-dimensional action.

(3) Verify that the NS5-brane is a solution of the ten-dimensional supergravity equations.

(4) Take the long-wavelength limit of the Horava–Witten theory. Write the Lagrangian in the ten-dimensional Einstein frame and verify that the gauge and gravitational couplings obey the relation appropriate to the heterotic string theory:

$$g_{\rm ym}^2 = 4\kappa^2 \alpha'^{-1}. \tag{28.71}$$

(5) Calculate the effective action of the matrix model at one loop in more detail. Verify that treated in Born approximation, this yields the correct graviton–graviton scattering matrix element for the eleven-dimensional theory. You may find the background field method helpful for this computation.

(6) Check that the configuration of Eq. (28.66) solves the field equations of IIB supergravity in the case of a single brane. You may want to use some of the available programs for evaluating the curvature. Verify that, in the Maldacena limit, the metric can be recast as in Eq. (28.30). If one requires that the curvature of the AdS space is small, check that the D-brane theory is strongly coupled. Discuss the problem of decoupling.

29

Large and warped extra dimensions

Considerations of the sort we encountered in the previous chapter have inspired two approaches to physics beyond the Standard Model: large extra dimensions (ADD) and warped spaces (Randall–Sundrum). In this chapter we will provide a brief introduction to each.

29.1 Large extra dimensions: the ADD proposal

In string theory, it is natural to imagine that the compactification scale is not too much different from the Planck scale. The size of the compact space is typically a modulus, and if it is stabilized, one might expect it be stabilized at a value not much different than one, in string (and therefore Planck) units. In terms of our general discussion of moduli stabilization, this is precisely what we would expect: once the radius becomes very large, any potential, perturbative or non-perturbative, tends to zero.

But if we are willing to discard this prejudice, an extraordinary possibility opens up. Perhaps the extra dimensions are not Planck size, but much larger, even macroscopic? Arkani-Hamed, Dimopoulos and Dvali realized that from an experimental point of view, the limits on the size of such large compact dimensions are surprisingly weak. Allowing the extra dimensions to be large totally reorients our thinking about the nature of couplings and scales in string theory (or any underlying fundamental theory). Such a viewpoint places the hierarchy problem in a whole different light, perhaps allowing entirely different solutions than technicolor or supersymmetry.

Branes are crucial to this picture. The observed gauge couplings are small, but not extremely small. But in Kaluza–Klein theory and in weakly coupled string theories, they are related to the underlying scales in a clear way. For example, in the heterotic string:

$$g_4^{-2} \cong g_{\rm s}^{-2} M_{\rm s}^6 R^6. \tag{29.1}$$

So if g_4 is fixed, as $R \to \infty$, $g_s \to \infty$. But even in a compactified theory, the gauge coupling on $D3$-branes is insensitive to the large volume. With more general branes, one has more intricate possibilities, depending on how the branes wrap the internal space. On the other hand, gravity becomes weak as R becomes large:

$$G_N = \frac{1}{M_p^2} = \frac{1}{M_p^8 R^6} = \frac{g_s^2}{\ell_s^8 R^6}. \tag{29.2}$$

Now if g_s is fixed and of order one, as $R \to \infty$, the Planck length tends to zero.

So how large might we imagine R could be? If we assume that R is macroscopic or nearly so, then on distance scales smaller than R, the force of gravity will be that appropriate to a higher-dimensional theory. In d space-time dimensions,

$$\text{Force}_g \sim \frac{1}{r^{d-2}}. \tag{29.3}$$

Experiments currently probe possible modifications of the gravitational force law on scales of order millimeters or somewhat smaller. (Since the proposal of large extra dimensions was put forward, these limits have been significantly improved.)

If the scale of the large extra dimensions is of order millimeters, how large is the fundamental scale? This depends on the number of dimensions that are actually large. If there are a large extra dimensions, any others being comparable in size to the fundamental scale,

$$M_p^2 = M_{\text{fund}}^{2+a} R^a, \tag{29.4}$$

or

$$M_{\text{fund}} = \left(M_p^2 R^{-a}\right)^{\frac{1}{2+a}} \quad R = M_p^{-1} (M_p/M_{\text{fund}})^{-(2+a)/a}. \tag{29.5}$$

A new viewpoint on the hierarchy problem arises by supposing that M_{fund} is close to the scale of weak interactions, say

$$M_{\text{fund}} \sim 1\,\text{TeV}. \tag{29.6}$$

Then we can use Eq. (29.5) to relate R to the Planck scale and the weak scale. For example, if $a = 2$, $R \approx 0.01$ cm! For larger a, it is smaller, but still dramatically large; for $a = 3$, for example, it is about 10^{-7} cm. But $a = 1$ would be, quite literally, astronomical in size, and is clearly ruled out by observations.

What is quite surprising is that it is difficult to rule out dimensions with size of order millimeters. Since the original proposal, there have been several experiments dedicated to improving the limits on deviations from Newtonian gravity at millimeter distances.

The possibility of large extra dimensions offers a different perspective on the hierarchy problem. The weak scale is fundamental; the issue is to understand why

the radius of the large dimensions is so large. One possibility which has been seriously considered is that there are some very large fluxes. For example, if H_{MN} is a two-form associated with a $U(1)$ gauge field, and Σ is some closed two-dimensional surface, we can have:

$$\int_\Sigma H_{MN} dx^M \wedge dx^N = N. \tag{29.7}$$

If the radius of the dimensions associated with Σ is large, then

$$H \sim \frac{N}{R^2}. \tag{29.8}$$

The potential, in turn, receives a contribution behaving as N^2/R^2. If there is also a (positive) cosmological constant,

$$V = \Lambda R^2 + \frac{N^2}{R^2} \tag{29.9}$$

and assuming that Λ is of order the fundamental scale,

$$R^4 \sim N^2 \ell_{\text{fund}}^4. \tag{29.10}$$

To obtain a sufficiently large radius in this way, then, requires an extremely large flux. There are some circumstances where such large pure numbers may not be required; supersymmetry and low dimensionality ($a = 2$) help.

For now, we will assume that somehow a large radius arises for dynamical reasons, and consider some of the other questions which, ultimately, such a picture raises.

(1) Proton decay: with no further assumptions about the theory, we would expect that baryon number violating operators would arise suppressed only by the TeV scale. It would then be necessary to suppress operators of very high dimension. One possible resolution of this problem is elaborate discrete symmetries. Another suggestion has been that the modes responsible for the different low-energy fermions might be very nearly orthogonal.

(2) Other flavor changing processes: for the same reason, flavor changing processes in weak interactions, processes like $\mu \to e + \gamma$, and the like pose a danger. One possible solution is a fundamental scale a few orders of magnitude larger than the weak scale. This raises the question of why the weak scale is small – the hierarchy problem again. Orthogonality of fermions, again, can help with many of these difficulties.

We turn, finally, to the phenomenology of large extra dimensions. Here there are exciting possibilities. If R is large, the Kaluza–Klein modes are very light. They are very weakly coupled, but there are lots of them and little energy is required for their production. So consider inclusive production of Kaluza–Klein particles in an accelerator. In terms of $G_N = \kappa^2/8\pi$ the amplitude for emission of a Kaluza–Klein

particle is proportional to κ. For any given mode, then, the cross section behaves as $\sigma_n \sim G_N E^2$, where the E^2 factor follows from dimensional analysis. We need to sum over n – equivalently, to integrate over a-dimensional phase space. As a crude estimate, we can treat the amplitude as constant, and cut off the integration at E, so

$$\sigma_{\rm tot} = R^a \int d^a k \, \sigma_k = G_N R^a E^{2+a}. \tag{29.11}$$

Recalling that $G_N = G_{\rm fund} R^{-a}$, we see that the tower of Kaluza–Klein particles couples like a $4 + a$-dimensional particle – i.e. at high energies, the extra dimensions are manifest! The cross section exhibits exactly the behavior with energy one expects in $4 + a$ dimensions.

The actual processes which might be observed in accelerators are quite distinctive. One would expect to see, for example, production of high-energy photons accompanied by missing energy,with the cross section showing a dramatic rise with energy. Such signatures have already been used (as of this writing) to set limits on such couplings.

The production of Kaluza–Klein particles in astrophysical environments can be used to set limits on extra dimensions as well. For example, in the case of two large dimensions and fundamental scale of order 1 TeV, we saw that the scale of the Kaluza–Klein excitations – the inverse of the radius of the extra dimensions – is of order 10^{-12} GeV, so such particles are easy to produce. Like axions, they could be readily produced in stars.

29.2 Warped spaces: the Randall–Sundrum proposal

Having entertained the possibility that some compact dimensions of space might be very large, one might wonder why the extra dimensions should be flat. The Horava–Witten theory provides a model. Taking the formulas of this theory literally, we have seen that if this theory describes nature, the eleventh dimension is quite large in fundamental units. The metric of this dimension is significantly distorted; we might say that it is warped. This is not surprising. The geometry is essentially one-dimensional. Green's functions for the fields grow linearly with distance. One of the appealing features of the Horava–Witten proposal is that the dimensions are just large enough that the distortion of the geometry is of order one.

Randall and Sundrum have made a more radical proposal: they argue that the warping might be enormous, and might account for the large hierarchy between the weak scale and the Planck scale. In the simplest version of their model, there is again one extra dimension; call its coordinate ϕ, $0 < \phi < \pi$. The model contains two branes, one at $\phi = 0$, one at $\phi = \pi$. The tensions of the two branes are taken to be equal and opposite. One imagines that the Standard Model fields propagate

on one brane, the "visible sector" brane, while some other, hidden sector fields propagate on the other. The action is:

$$S = S_{\rm grav} + S_{\rm vis} + S_{\rm hid}. \tag{29.12}$$

The bulk gravitational action, $S_{\rm grav}$, includes a cosmological constant term:

$$S_{\rm grav} = \int d^4x \int d\phi \sqrt{-G}[-\Lambda + 2M^3\mathcal{R}], \tag{29.13}$$

where M is the five-dimensional Planck mass. The brane actions are:

$$S_{\rm vis} = \int d^4x \sqrt{-g_{\rm vis}}[\mathcal{L}_{\rm vis} - \Lambda_{\rm vis}] \quad S_{\rm hid} = \int d^4x \sqrt{-g_{\rm hid}}[\mathcal{L}_{\rm hid} - \Lambda_{\rm hid}]. \tag{29.14}$$

Here we have separated off a brane tension term on each brane; we have also distinguished the bulk five-dimensional metric, G_{MN}, from the metrics on each of the branes, $g_{\mu\nu}$. This has the structure of a gravitational problem in five dimensions, with δ-function sources at $\phi = 0, \pi$. Einstein's equations are:

$$\sqrt{-G}\left(R_{MN} - \frac{1}{2}G_{MN}R\right) = -\frac{1}{4M^3}\Big[\Lambda\sqrt{-G}G_{MN} + \Lambda_{\rm vis}\sqrt{-g_{\rm vis}}g^{\rm vis}_{\mu\nu}\delta^\mu_M\delta^\nu_N\delta(\phi-\pi) + \Lambda_{\rm hid}\sqrt{-g_{\rm hid}}g^{\rm hid}_{\mu\nu}\delta^\mu_M\delta^\nu_N\delta(\phi)\Big]. \tag{29.15}$$

Now one makes an ansatz for the metric, which leads to warping:

$$ds^2 = e^{-2\sigma(\phi)}\eta_{\mu\nu}dx^\mu dx^\nu + r_c^2 d\phi^2. \tag{29.16}$$

Plugging in the ansatz, Eq. (29.16), one obtains equations for σ:

$$\frac{6\sigma'^2}{r_c^2} = \frac{-\Lambda}{4M^3} \quad \frac{3\sigma''}{r_c^2} = \frac{\Lambda_{\rm hid}}{4M^3 r_c}\delta(\phi) + \frac{\Lambda_{\rm vis}}{4M^3 r_c}\delta(\phi - \pi). \tag{29.17}$$

This is solved by:

$$\sigma = r_c|\phi|\sqrt{-\frac{\Lambda}{24M^3}} \tag{29.18}$$

provided that the following conditions on the Λs hold:

$$\Lambda_{\rm hid} = \Lambda_{\rm vis} = 24M^3k \quad \Lambda = -24M^3k^3. \tag{29.19}$$

In this case, the metric varies exponentially rapidly. Note that r_c does not need to be terribly large in order that one obtain an enormous hierarchy. One might worry, though, about the identification of the graviton. It turns out that the metric has zero modes:

$$ds^2 = e^{-2kr_c|\phi|}[\eta_{\mu\nu} + \tilde{h}_{\mu\nu}(x)dx^\mu dx^\nu + T^2(x)d\phi^2], \tag{29.20}$$

where T^2 represents a variation of r_c, and is usually referred to as the radion, and $\tilde{h}_{\mu\nu}$ is the four-dimensional metric. If one substitutes in the action, one finds:

$$S = \int d^4x \int d\phi 2M^3 r_c e^{-2kr_c|\phi|} \sqrt{-\tilde{g}}\tilde{R}. \tag{29.21}$$

From this we can read off the effective Planck mass:

$$M_{\rm p}^2 = M^3 r_c \int d\phi e^{-2kr_c|\phi|} = \frac{M^3}{k}\left[1 - e^{-2kr_c}\right]. \tag{29.22}$$

So the four-dimensional Planck scale is comparable to the fundamental five-dimensional scale.

To see that the physical masses on the visible brane are small, consider the visible sector action for a scalar particle:

$$S_{\rm vis} = \int d^4x\sqrt{-g}e^{-4kr_c\pi}\left[\tilde{g}_{\mu\nu}e^{2kr_c\pi}|D_\mu\phi|^2 - \lambda\left(|\phi|^2 - v_0^2\right)^2\right]. \tag{29.23}$$

Rescaling $\phi \to e^{kr_c\pi}\phi$, we have:

$$S_{\rm vis} = \int d^4x\sqrt{-g}\left[\tilde{g}_{\mu\nu}|D_\mu\phi|^2 - \lambda\left(|\phi|^2 - e^{-2kr_c\pi}v_0^2\right)^2\right], \tag{29.24}$$

so the scale is indeed exponentially smaller than the scale on the other brane.

There are many questions one can ask about this structure.

(1) How robust is this sort of localization of gravity?
(1) How do higher excitations, e.g. bulk fields, interact with the fields on the brane? Is the hierarchy stable? (The answer is yes.)
(3) Does this sort of warping arise in string theory? Again, the answer is yes, though the details look different.
(4) As in the case of large extra dimensions, if this picture makes sense, there are many excitations on the branes. Higher-dimension operators are suppressed only by the TeV scale. As there, one has to ask: how does one understand conservation of baryon number? Other flavor changing processes? Neutrino masses? Precision electroweak physics? Answers have been put forward to all of these questions, but they remain suitable subjects for research.
(5) Assuming the above problems are resolved, what are the experimental signals for such warping? As in the case of large extra dimensions, one wants to focus on the additional degrees of freedom associated with bulk fields and the brane. In this case, unlike the case of large extra dimensions, the Kaluza–Klein states are not dense. Instead, the low-lying states have masses and spacings of order the TeV scale. Their couplings are not of gravitational strength, but instead scaled by inverse powers of the scale of the visible sector brane.

Finally, there are other variants of the Randall–Sundrum proposal which have been put forward. Perhaps the most interesting is one in which space is not

compactified at all, but simply warped, with gravity localized on the visible brane. These ideas suggest a rich set of possibilities for what might underlie a quantum theory of gravity. Some of these features – the exponential warping of the metric, in particular – have been observed in string theory, but many, at least to date, have not. This is a potentially important area for further research.

Suggested reading

The original paper of Arkani-Hamed *et al*. (1999) is quite clear and comprehensive, as is the paper of Randall and Sundrum (1999). The phenomenology of the Randall–Sundrum models is explored by Davoudiasl *et al*. (2000).

Exercise

(1) Verify the Randall–Sundrum solution of Eq. (29.15).

30

Coda: Where are We Headed?

As this book is being completed, the Large Hadron Collider (LHC) at CERN, and its two large detectors, ATLAS and CMS, are nearing completion. The center of mass energy at this machine will be large, about 14 TeV. The center of mass energies of the partons – the quarks and gluons – within the colliding protons will be larger than 1 TeV. The luminosity will also be very large. As a result, if almost any of the ideas we have described for understanding the hierarchy problem in Part 1 of this book are correct, evidence should appear within a few years. For example, if the hypothesis of low-energy supersymmetry is correct, we should see events with large amounts of missing energy, and signatures such as multiple leptons. Large extra dimensions should be associated with rapid growth of cross sections for various processes, again with missing energy; the warped spaces suggested by Randall and Sundrum should be associated with the appearance of massive resonances. Technicolor, similarly, should lead to broad resonances. Assuming some underlying technicolor model can satisfy constraints from flavor physics and precision electroweak measurements, one might expect to find some number of light (compared with 1 TeV), pseudo-Goldstone bosons, many with gauge quantum numbers. If any of these phenomena occur, distinguishing among them in the complicated environment of a hadron machine will be challenging. It is conceivable that there will be competing explanations, and that choosing between them will require a very high-energy electron–positron colliding beam machine. Such a machine is under consideration by a consortium of nations, and is referred to as the *International Linear Collider*, or ILC. Hopefully in later editions of this book, it will be possible to focus on real experimental results, rather than a range of theoretical speculation.

Before the data rolls in, we might hope to select among these possibilities, or perhaps discover some crucial idea – and possible set of phenomena – that we are missing. Here, string theory might help. Many of the ideas for Beyond the Standard Model Physics require phenomena which can only be understood within

a quantum theory of gravity. This is certainly true of large extra dimensions or Randall–Sundrum.

We have discussed many aspects of string theory, but as far as the world about us and the experiments which explore it, we have left things in a rather unsettled state. We have seen that string theories have many of the features we might hope for from nature. We have exhibited ground states – more precisely, approximate moduli spaces – with many of the features of the Standard Model: the observed gauge groups, repetitive generations of quarks and leptons, calculable gauge and Yukawa couplings, and more. But it is not clear how to make sharp predictions. There are vast numbers of moduli spaces with the wrong features: the wrong number of dimensions, too much supersymmetry, the wrong gauge group and matter content, and we have not offered a dynamical mechanism or principle which might select among them. Not only are there discrete choices, but there are continuous ones as well, associated with the moduli. We have seen how potentials for the moduli arise, but we have not offered any idea for how stable or metastable vacua might arise, other than to note that such states will typically lie at strong coupling, where they are inaccessible to analysis. Note that our discussion of strong–weak coupling duality, by itself, does not help with this problem; our general arguments show that one cannot find stable vacua at arbitrarily weak coupling in any description.

For a long time, string theorists hoped for some *deus ex machina* which might resolve this conundrum. Some have imagined that one would simply find some new type of string model or construction which would not suffer from these difficulties, perhaps leading uniquely to the Standard Model at low energies. Developments in string duality have suggested an alternative picture: there might exist a vast number of isolated, stable or metastable states of the theory, with little or no supersymmetry. These states seem to have a distribution of values of couplings, mass scales and cosmological constant. There may be an exponentially large or even infinite number of them.

To understand how these come about, we return, again, to the IIB theories. In these theories, there are two three-form gauge fields, one arising from the NS–NS sector, the other in the R–R sector; denote these H and F. If one compactifies on a Calabi–Yau space, it is possible to have non-trivial backgrounds for these fields. One can define three-form fluxes by integrating these over non-trivial three cycles. The number of such cycles, we have learned, is $h_{2,1}$. The fluxes are quantized, by an argument identical to Dirac's. There are also constraints on the values of fluxes, resulting from absence of anomalies (violation of Gauss's law). In the presence of fluxes, we might expect the system to have a non-trivial energy; this energy is a function of the values of the moduli, so there is a potential for the moduli. This potential can be described in terms of a superpotential. It turns out that there is a

simple formula for the superpotential:

$$W = \int G_3 \wedge \Omega, \tag{30.1}$$

where $G_3 = F_3 - \tau H_3$ and ω is the covariantly constant three-form. One can obtain explicit formulas in the case of compactification of the IIB theory on a torus. This expression depends on the complex structure moduli, z, but not on the Kahler moduli, ρ.

The equations $D_z W = 0$ generically fix the complex structure moduli and the dilaton but not the Kahler moduli. But the superpotential of Eq. (30.1) is not exact, and can receive non-perturbative corrections. These can lead to a potential for the Kahler moduli. Logically, we can organize the analysis by first integrating out the complex structure moduli and the dilaton, leaving a superpotential for the Kahler modulus, ρ, of the form:

$$W_\rho = W_0 + e^{-c\rho}. \tag{30.2}$$

The Kahler potential, on the other hand, for large ρ and weak coupling should not be too different than its tree-level form:

$$K = -3\ln(\rho + \rho^*). \tag{30.3}$$

If W_0 is small, it is not hard to see that the potential has a supersymmetric stationary point at

$$\rho \approx -c\ln(W_0). \tag{30.4}$$

But why should W_0 be small? The answer is that, in general, it isn't, but there are many possible choices of flux, and so there are many different possible states with different values of W_0. In this construction, the number of possible flux choices which permit a supersymmetric low-energy theory for ρ is finite but extremely large. One can understand this by thinking of a b-dimensional vector of integers, $\vec{N}$, representing the fluxes, constrained by $|\vec{N}|^2 < L^2$. The number of possible flux choices is then the volume of a b-sphere of radius L, or of order

$$\mathcal{N} \approx L^b. \tag{30.5}$$

In interesting Calabi–Yau compactifications, L and b can be of order 100s.

If we suppose that W_0 is distributed more or less randomly among these states, we might expect that the probability of finding W_0 at small W_0 would be roughly uniform with W_0 as a complex variable, i.e. $\int d^2 W_0 P(W_0)$, with $P(W_0)$ constant. Numerical studies suggest that this is indeed the case.

Other quantities will also vary across the landscape. One can get gauge groups, for example, by including branes in these configurations. There will be, then, a

distribution of groups and couplings. Perhaps not surprisingly, the distribution of *gauge couplings*, at small gauge coupling, is typically flat in g^2. One might expect to find supersymmetry broken dynamically or through brane configurations. There is also a vast array of states – perhaps infinite, where supersymmetry is broken already by the classical superpotential for the complex structure moduli, i.e. one cannot solve the set of equations $D_z W = 0$.

This is compatible with many of our ideas for thinking about the hierarchy problem. In supersymmetry, technicolor, and the Randall–Sundrum approaches, one argues that it is reasonable to have exponentially small scales, because it is reasonable to think that couplings should be small, but not much smaller than one. Why this should be so is unclear. If there is a unique underlying theory, for example, the couplings are whatever they are. But within this landscape, there is a distribution, which would seem to motivate just the sorts of ideas we have explored.

Much work is going into mapping the features of this landscape. Many types of states have been explored and their statistics at least partially understood, including states with and without supersymmetry, with various gauge groups, and the like. From these studies, it seems plausible that there are some – possibly many – states in the landscape with the low-energy features of the Standard Model. This naturally raises several related questions.

(1) Are typical states with the features of the Standard Model characterized by things we might hope to measure, such as large extra dimensions or supersymmetry?
(2) Why do we find ourselves in a state which looks like ours does?

If one can answer these questions, one might hope to make real predictions. It is, at this writing, too early to answer the first, though one can enumerate some of the issues. For example, if there are comparable numbers of supersymmetric and non-supersymmetric states, it might be that there are more supersymmetric states with a large hierarchy of states than non-supersymmetric ones. If there are vastly more non-supersymmetric states, it might be that technicolor or warping are the most common ways of obtaining large hierarchies.

If the landscape really exists with the sorts of properties we are attributing to it, it seems likely that the answer to the second question is, in part, *environmental*. The issue is posed most starkly by the cosmological constant. The distribution of cosmological constants is not likely to peak at the very tiny but non-zero observed value. On the other hand, Weinberg, and also Banks, pointed out long ago that if the cosmological constant were much larger than observed (perhaps a factor of 10), galaxies would not form. The problem is that the exponential expansion of the universe would begin before the perturbations from inflation became non-linear. Given that structure formed when the universe was about $1/20$ of its present age, a cosmological constant far larger than the present energy density would be

problematic. This argument predicted a cosmological constant in the observed range.

The idea that some quantities are determined by environmental considerations, or *anthropic* conditions, is quite controversial. At one level, one can ask what would happen if several quantities, such as the parameters of the inflationary potential, were allowed to vary. At another, this argument requires that the universe, somehow in its history, sampled all of the different possible states. Within present ideas about inflation, it is not clear whether this is possible. Finally, even granting the first two points, this argument rests on the requirement of observers to determine at least some of the laws of nature. Some physicists argue that this is no different than saying that we find ourselves on a planet at a certain distance from a star because, otherwise, the planet could not support life; they point out that through much of human history, the Earth–Sun distance was viewed as fundamental. But many physicists find this troubling. At this writing, however, it is fair to say that this is by far the most plausible explanation we have of the value of Λ.

Hopefully, the reader who has worked patiently through this book, has been left with challenging questions, and the skills to attack them. As the LHC turns on, as further progress is made in understanding how quantum gravity might be reconciled with much lower-energy physics, we may be at the threshold of understanding what lies beyond the Standard Model.

Suggested reading

Weinberg's argument for the cosmological constant appears in his 1989 paper. The landscape picture we have described here was first put forward by Bousso and Polchinski (2000) and Feng *et al*. (2001), and formulated most sharply in string theory by Kachru *et al*. (2003). Some statistical studies, which significantly elucidate the possible meaning of the landscape, appear in Ashok and Douglas (2004).

Part 4

The appendices

Appendix A

Two-component spinors

The massless Dirac equation simplifies dramatically in the case that the fermion mass is zero. The equation

$$\not{D}\psi = 0 \tag{A1}$$

has the feature that if ψ is as solution, so is $\gamma_5\psi$:

$$\not{D}(\gamma_5) = 0. \tag{A2}$$

The matrices

$$P_\pm = \frac{1}{2}(1 \pm \gamma_5) \tag{A3}$$

are projectors,

$$P_\pm^2 = P_\pm \quad P_+P_- = P_-P_+ = 0. \tag{A4}$$

To understand the physical significance of these projectors, it is convenient to use a particular basis for the Dirac matrices, often called the chiral or Weyl basis:

$$\gamma^\mu = \begin{pmatrix} 0 & \sigma^\mu \\ \bar{\sigma}^\mu & 0 \end{pmatrix}, \tag{A5}$$

where

$$\sigma^\mu = (1, \vec{\sigma}) \quad \bar{\sigma}^\mu = (1, -\vec{\sigma}). \tag{A6}$$

In this basis,

$$\gamma_5 = i\gamma^0\gamma^1\gamma^2\gamma^3 = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{A7}$$

so

$$P_+ = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \quad P_- = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}. \tag{A8}$$

We will adopt some notation, following the text by Wess and Bagger:

$$\psi = \begin{pmatrix} \chi_\alpha \\ \phi^{*\dot{\alpha}} \end{pmatrix}. \tag{A9}$$

Correspondingly, we label the indices on the matrices σ^μ and $\bar{\sigma}^\mu$ as

$$\sigma^\mu = \sigma^\mu_{\alpha\dot{\alpha}} \quad \bar{\sigma}^\mu = \bar{\sigma}^{\mu\dot{\beta}\beta}. \tag{A10}$$

This allows us to match upstairs and downstairs indices, and will prove quite useful. The Dirac equation now becomes:

$$i\sigma^\mu_{\alpha\dot{\alpha}}\partial_\mu\phi^{*\dot{\alpha}} = 0 \quad i\bar{\sigma}^{\mu\dot{\alpha}\alpha}\partial_\mu\chi_\alpha = 0. \tag{A11}$$

Note that χ and ϕ^* are equivalent representations of the Lorentz group; χ and ϕ obey identical equations. We may proceed by complex conjugating the second equation in Eq. (A11), and noting $\sigma_2\sigma^{\mu *}\sigma_2 = \bar{\sigma}^\mu$.

Before discussing this identification in terms of representations of the Lorentz group, it is helpful to introduce some further notation. First, we define complex conjugation to change dotted to undotted indices. So, for example,

$$\phi^{*\dot{\alpha}} = (\phi^\alpha)^*. \tag{A12}$$

Then we define the antisymmetric matrices $\epsilon_{\alpha\beta}$ and $\epsilon^{\alpha\beta}$ by:

$$\epsilon^{12} = 1 = -\epsilon^{21} \quad \epsilon_{\alpha\beta} = -\epsilon^{\alpha\beta}. \tag{A13}$$

The matrices with dotted indices are defined identically. Note that, with upstairs indices, $\epsilon = i\sigma_2$, $\epsilon_{\alpha\beta}\epsilon^{\beta\gamma} = \delta^\gamma_\alpha$. We can use these matrices to raise and lower indices on spinors. Define $\phi_\alpha = \epsilon_{\alpha\beta}\phi^\beta$, and similar for dotted indices. So

$$\phi_\alpha = \epsilon_{\alpha\beta}(\phi^{*\dot{\beta}})^*. \tag{A14}$$

Finally, we will define complex conjugation of a product of spinors to invert the order of factors, so, for example, $(\chi_\alpha\phi_\beta)^* = \phi^*_{\dot{\beta}}\chi^*_{\dot{\alpha}}$.

With this in hand, the reader should check that the action for our original four-component spinor is:

$$\begin{aligned} S = \int d^4x\mathcal{L} &= \int d^4x \left(i\chi^*_{\dot{\alpha}}\bar{\sigma}^{\mu\dot{\alpha}\alpha}\partial_\mu\chi_\alpha + i\phi^\alpha\sigma^\mu_{\alpha\dot{\alpha}}\partial_\mu\phi^{*\dot{\alpha}}\right) \\ &= \int d^4x\mathcal{L} = \int d^4x \left(i\chi^\alpha\sigma^\mu_{\alpha\dot{\alpha}}\partial_\mu\chi^{*\dot{\alpha}} + i\phi^\alpha\sigma^\mu_{\alpha\dot{\alpha}}\partial_\mu\phi^{*\dot{\alpha}}\right). \end{aligned} \tag{A15}$$

At the level of Lorentz-invariant Lagrangians or equations of motion, there is *only one* irreducible representation of the Lorentz algebra for massless fermions.

Two-component fermions have definite helicity. For a single-particle state with momentum $\vec{p} = p\hat{z}$, the Dirac equation reads:

$$p(1 \pm \sigma_z)\phi = 0. \tag{A16}$$

Similarly, the reader should check that the antiparticle has the opposite helicity.

It is instructive to describe quantum electrodynamics with a massive electron in two-component language. Write

$$\psi = \begin{pmatrix} e \\ \bar{e}^* \end{pmatrix}. \tag{A17}$$

In the Lagrangian, we need to replace ∂_μ with the covariant derivative, D_μ. Note that e contains annihilation operators for the left-handed electron, and creation operators for the corresponding antiparticle. Note also that $\bar{e}$ contains annihilation operators for a particle with the opposite helicity and charge of e, and $\bar{e}^*$, and creation operators for the corresponding antiparticle.

The mass term, $m\bar{\psi}\psi$, becomes:

$$m\bar{\psi}\psi = me^{\alpha}\bar{e}_{\alpha} + me^{*}_{\dot{\alpha}}\bar{e}^{*\dot{\alpha}}. \tag{A18}$$

Again, note that both terms preserve electric charge. Note also that the equations of motion now couple e and $\bar{e}$.

It is helpful to introduce one last piece of notation. Call

$$\psi\chi = \psi^{\alpha}\chi_{\alpha} = -\psi_{\alpha}\chi^{\alpha} = \chi^{\alpha}\psi_{\alpha} = \chi\psi. \tag{A19}$$

Similarly,

$$\psi^{*}\chi^{*} = \psi^{*}_{\dot{\alpha}}\chi^{*\dot{\alpha}} = -\psi^{*\dot{\alpha}}\chi^{*}_{\dot{\alpha}} = \chi^{*}_{\dot{\alpha}}\psi^{*\dot{\alpha}} = \chi^{*}\psi^{*}. \tag{A20}$$

Finally, note that with these definitions,

$$(\chi\psi)^{*} = \chi^{*}\psi^{*}. \tag{A21}$$

Appendix B

Goldstone's theorem and the pi mesons

It is easy to prove Goldstone's theorem for theories with fundamental scalar fields. But the theorem is more general, and some of its most interesting applications are in theories without fundamental scalars. We can illustrate this with QCD. In the limit that there are two massless quarks (i.e. in the limit that we neglect the mass of the u and d quarks), we can write the QCD Lagrangian in terms of spinors

$$\Psi = \begin{pmatrix} u \\ d \end{pmatrix} \tag{B1}$$

as

$$\mathcal{L} = \bar{\psi} i \gamma^\mu D_\mu \psi - \frac{1}{4} F_{\mu\nu}^2. \tag{B2}$$

This Lagrangian has symmetries:

$$\Psi \to e^{i\omega^a \frac{\tau^a}{2}} \Psi \qquad \Psi \to e^{i\omega^a \frac{\tau^a}{2} \gamma^5} \Psi \tag{B3}$$

(τ^a are the Pauli matrices). In the limit that two quarks are massless, QCD is thus said to have the symmetry $SU(2)_{\mathrm{L}} \times SU(2)_{\mathrm{R}}$.

So writing a general four-component fermion as

$$q = \begin{pmatrix} q \\ \bar{q}^* \end{pmatrix}, \tag{B4}$$

the Lagrangian has the form:

$$\mathcal{L} = i \Psi \sigma^\mu D_\mu \Psi^* + i \bar{\Psi} \sigma^\mu D_\mu \bar{\Psi}^*. \tag{B5}$$

In this form, we have two separate symmetries:

$$\Psi \to e^{i\omega_{\mathrm{L}}^a \frac{\tau^a}{2}} \Psi \qquad \bar{\Psi} \to e^{i\omega_{\mathrm{R}}^a \frac{\tau^a}{2}} \bar{\Psi}. \tag{B6}$$

Written in this way, it is clear why the symmetry is called $SU(2)_{\mathrm{L}} \times SU(2)_{\mathrm{R}}$.

Now it is believed that in QCD, the operator $\overline{\Psi}\Psi$ has a non-zero vacuum expectation value, i.e.

$$\langle \overline{\Psi}\Psi \rangle \approx (0.3\mathrm{GeV})^3 \delta_{ff'}. \tag{B7}$$

This is in four-component language; in two-component language this becomes:

$$\langle \bar{\Psi}_f \Psi_{f'} + \bar{\Psi}_f^* \Psi_{f'}^* \rangle \neq 0. \tag{B8}$$

This leaves ordinary isospin, the transformation, in four-component language, without the γ_5, or, in two-component language, with $\omega_L^a = -\omega_R^a$, unbroken.

But there are three broken symmetries. Correspondingly, we expect that there are three Goldstone bosons. To prove this, write

$$\mathcal{O} = \bar{\Psi}\Psi; \quad \mathcal{O}^a = \bar{\Psi}\gamma^5 \frac{\tau^a}{2}\Psi. \tag{B9}$$

Under an infinitesimal transformation,

$$\delta\mathcal{O} = 2i\omega^a\mathcal{O}^a; \quad \delta\mathcal{O}^a = i\omega^a\mathcal{O}. \tag{B10}$$

In the quantum theory, these becomes the commutation relations:

$$[Q^a, \mathcal{O}] = 2i\mathcal{O}^a; \quad [Q^a, \mathcal{O}^b] = i\delta^{ab}\mathcal{O}. \tag{B11}$$

Q^a is the integral of the time component of a current. To see that there must be a massless particle, we study

$$0 = \int d^4x \partial_\mu [\langle\Omega|T(j^{\mu a}(x)\mathcal{O}^b(0))|\Omega\rangle e^{-ip\cdot x}] \tag{B12}$$

(this just follows because the integral of a total derivative is zero). We can evaluate the right-hand side, carefully writing out the time-ordered product in terms of θ functions, and noting that ∂_0 on the θ functions gives δ-functions:

$$0 = \int d^4x \langle\Omega|[j^{oa}(x), \mathcal{O}^b(0)]\delta(x^0)|\Omega\rangle e^{-ip\cdot x} - ip_\mu \times \int d^4x \langle\Omega|T(j^{\mu a}(x)\mathcal{O}^b(0))|\Omega\rangle. \tag{B13}$$

Now consider the limit $p^\mu = 0$. The first term on the right-hand side becomes the matrix element of $[Q^a, \mathcal{O}^b(0)] = \mathcal{O}(0)$. This is non-zero. The second term must be singular, then, if the equation is to hold. This singularity, as we will now show, requires the presence of a massless particle. For this we use the spectral representation of Green's function. In general, a pole can arise at zero momentum only from a massless particle. To understand this singularity we introduce a complete set of states, and, say for $x^0 > 0$, write it as

$$\sum_\lambda \int \frac{d^3p}{2E_p(\lambda)} \langle\Omega|j^{\mu a}(x)|\lambda_p\rangle\langle\lambda_p|\mathcal{O}^b(0)|\Omega\rangle. \tag{B14}$$

In the sum, we can separate the term from the massless particle. Call this particle π^b. On Lorentz-invariance grounds,

$$\langle\Omega|j^{\mu a}|\pi^b(p)\rangle = f_\pi p^\mu \delta^{ab}. \tag{B15}$$

Call

$$\langle\lambda_q|\mathcal{O}^a(x)|\pi^b(p)\rangle = Z\delta^{ab}, e^{-ip\cdot x} \tag{B16}$$

Considering the other time ordering, we obtain for the left-hand side a massless scalar propagator, i/p^2, multiplied by $Zf_\pi p^\mu$, so the equation is now consistent:

$$\langle\bar{\Psi}\Psi\rangle = \frac{p^2}{p^2} f_\pi Z. \tag{B17}$$

It is easy to see that in this form, Goldstone's theorem generalizes to any theory without fundamental scalars in which a global symmetry is spontaneously broken.

Returning to QCD, what about the fact that the quarks are massive? The quark mass terms break explicitly the symmetries. But if these masses are small, we should be able to

think of the potential as "tilted." This gives rise to a small mass for the pions. We can compute these by studying, again, correlation functions of derivatives of currents. A simpler procedure is to consider the symmetry-breaking terms in the Lagrangian:

$$\mathcal{L}_{\text{sb}} = \bar{\Psi} M \Psi, \tag{B18}$$

where M is the quark mass matrix,

$$M = \begin{pmatrix} m_u & 0 \\ 0 & m_d \end{pmatrix}. \tag{B19}$$

Since the π mesons are, by assumption, light, we can focus on these. If we have a non-zero pion field, we can think of the fermions as being given by:

$$\Psi = e^{i \frac{\pi^a}{f_\pi} \gamma^5 \frac{\tau^a}{2}} \Psi. \tag{B20}$$

In other words, the pion fields are like symmetry transformations of the vacuum (and everything else).

Now assume that there is an "effective interaction" for the pions containing kinetic terms $(1/2)(\partial_\mu \pi^a)^2$. Taking the form above for Ψ, the pions get a potential from the fermion mass terms. To work out this potential, one plugs this form for the fermions into the Lagrangian and replaces the fermions by their vacuum expectation value. This gives that

$$V(\pi) = \langle \bar{q} q \rangle \text{Tr}\left(e^{i\omega^a \gamma_5 \tau^a} M\right), \tag{B21}$$

one can expand to second order in the pion fields, obtaining:

$$m_\pi^2 f_\pi^2 = (m_u + m_d)\langle \bar{q} q \rangle. \tag{B22}$$

Exercises

(1) Verify Eq. (B13).
(2) Derive Eq. (B22), known as the Gell-Mann–Oakes–Renner formula.

Appendix C

Some practice with the path integral in field theory

The path integral is extremely useful, both in field theory and in string theory. This appendix provides a brief review of path integration, and some applications. Many of the examples are drawn from finite-temperature field theory. These are instructive since one can easily write very explicit expressions. They are also useful for understanding the high-temperature universe, and are closely connected to computations which arise in compactified theories.

C.1 Path integral review

Feynman gave an alternative formulation of quantum mechanics, in which one calculates amplitudes by summing over possible trajectories of a system, weighting by $e^{iS/\hbar}$, where S is the classical action of the trajectory. For a particle, the path integral is:

$$Z = \int [dx] e^{iS/\hbar}. \tag{C1}$$

Here $\int [dx]$ is an instruction to sum over all possible paths of the particle.

This generalizes immediately to field theory, where surprisingly it is often more useful:

$$Z = \int [d\phi] e^{iS}. \tag{C2}$$

For a single field, ϕ, it is useful to introduce sources, $J(x)$, and to define

$$Z[J] = \int [d\phi] e^{i\int d^4x (\frac{1}{2}(\partial\phi)^2 - V(\phi) + J\phi)}. \tag{C3}$$

Green's functions of ϕ can then be obtained by functional differentiation of Z with respect to J:

$$T\langle\phi(x_1)\dots\phi(x_n)\rangle = \frac{\delta}{i\delta J(x_1)} \cdots \frac{\delta}{i\delta J(x_n)} Z[J]. \tag{C4}$$

For free fields, the integral can be performed by completing the squares. Writing the action as:

$$S_{\text{free}} = \int d^4x \left(\frac{1}{2}\phi(x) D^{-1} \phi(x) + \phi(x) J(x)\right) \tag{C5}$$

with

$$D^{-1} = \partial^2 - m^2 = p^2 - m^2. \tag{C6}$$

We can complete the squares in the action:

$$S_{\text{free}} = \int d^4x \left(\frac{1}{2}\phi(x) + \int d^4y J(y)D(y,x)\right) D^{-1} \left(\phi(x) + \int d^4z J(z)D(z,x)\right) - \int d^4x d^4y J(x)D(x,y)J(y). \tag{C7}$$

Now in the free field functional integral, one can shift the ϕ integral, obtaining:

$$Z_0[J] = \Delta e^{\frac{-i}{2}\int d^4x d^4y J(x)D(x,y)J(y)}. \tag{C8}$$

Here Δ is the free field functional integral at $J = 0$. It is the (square root) of the functional determinant of the operator D. D itself is the propagator of the scalar. This expression can then be used to develop perturbation theory. For example, with a $(\lambda/4!)\phi^4$ interaction, we can write:

$$Z[J] = \exp\left(i \int d^4x \frac{\lambda}{4!}\left(\frac{\delta}{i\delta J(x)}\right)^4\right) Z_0[J]. \tag{C9}$$

Working out the terms in the power series reproduces precisely the Feynman diagram expansion.

This has generalizations to non-Abelian gauge theories, with both unbroken and broken symmetries, which we will discuss in the text. We will also find it useful for addressing other questions.

C.2 Finite-temperature field theory

As an application of path integral methods and because of its importance in cosmology, we consider at some length the problem of field theory at finite temperature.

In statistical mechanics, one is interested in the partition function,

$$Z[\beta] = \text{Tr} e^{-\beta H}. \tag{C10}$$

For a quantum mechanical system, in contact with a heat bath, this is:

$$Z[\beta] = \sum_n \langle n|e^{-\beta E n}|n\rangle, \tag{C11}$$

where n label the energy eigenstates.

For a harmonic oscillator (unit mass), $H = [(p^2/2) + (\omega^2/2)]x^2$, and the partition function is:

$$e^{-\beta F} = \sum_n e^{-\beta\omega(n+\frac{1}{2})} = e^{-\omega\beta/2}\frac{1}{1 - e^{-\beta\omega}}. \tag{C12}$$

Now we can think of

$$\langle x|e^{-\beta H}|x\rangle \tag{C13}$$

as the amplitude that starting at x one ends up at x after propagating through an imaginary time $-i\beta$. This can be represented as a path integral:

$$\langle x|e^{-\beta H}|x\rangle = \int_{x(0)=x(\beta)=x} [dx] e^{-\int_0^\beta dt L_{\mathrm{E}}}, \tag{C14}$$

where L_{E} is the Euclidean Lagrangian,

$$L_{\mathrm{E}} = \left(\frac{dx}{dt}\right)^2 + \frac{1}{2}\omega^2 x^2 \tag{C15}$$

(note the signs here!). The partition function is now

$$Z[\beta] = \int_{x(0)=x(\beta)=x_0}^{dx_0} [dx] e^{-\int_0^\beta dt L_{\mathrm{E}}}, \tag{C16}$$

i.e. we integrate over the possible values of x at $t = 0$ in order to take the trace. This is the problem of a periodic box in the time direction. For this simple system with one degree of freedom, we can write:

$$x(t) = \sum_n \frac{1}{\sqrt{T}} a_n e^{-\frac{2\pi i n}{\beta} t}. \tag{C17}$$

We will simplify the problem slightly by taking $x(t)$ to be complex (you can think of this simply as an isotropic harmonic oscillator in two dimensions). The action of this configuration is:

$$S = \sum_{n=-\infty}^{\infty} \frac{1}{2}\left(\omega_n^2 + \omega^2\right)|a_n|^2. \tag{C18}$$

The path integral is now:

$$Z[\beta] = \prod \int da_n da_n^* e^{-S_{\mathrm{E}}}. \tag{C19}$$

The integrals are just Gaussian integrals. For a complex variable, z, we have

$$\int d^2 z e^{-a|z|^2} = \frac{\pi}{a} \tag{C20}$$

so we have the result for Z:

$$Z[\beta] = \prod \frac{1}{\omega^2 + \omega_n^2}, \tag{C21}$$

where $\omega_n = (2\pi n)/T$.

Now before trying to evaluate this product, it is useful to pause and note that this can be expressed in terms of the determinant of a matrix. Quite generally, Gaussian path integrals take the form of (inverse) determinants. In this case, if we call $\mathcal{M}$ the differential operator:

$$\mathcal{M} = \frac{1}{2}\left(-\frac{d^2}{dt^2} + \omega^2\right) \tag{C22}$$

its eigenfunctions are just $e^{i\omega_n t}$, with eigenvalues $\omega_n^2 + \omega^2$. So Z is just the inverse determinant of $\mathcal{M}$. Had we worked with only one real coordinate, we would have obtained the square root of the inverse determinant.

The determinant of an infinite matrix may seem a daunting object, but there are some tricks that permit evaluation in many cases. The first thing is to write the determinant as a sum, by taking a logarithm. In general,

$$\det M = e^{\,\mathrm{Tr}\ln(M)} \tag{C23}$$

(to see this, diagonalize M). It is easier to evaluate derivatives of the determinant rather than the determinant itself. We can derive a very useful formula for the derivative of a determinant by writing:

$$\begin{aligned}\det(M+\delta M) &= e^{\,\mathrm{Tr}\ln(M+\delta M)} = e^{(\mathrm{Tr}\ln(M)+\ln(1+M^{-1}\delta M))} \\ &= e^{\,\mathrm{Tr}\ln(M)} e^{\,\mathrm{Tr}M^{-1}\delta M} \approx \det M(1+\ \mathrm{Tr}M^{-1}\delta M).\end{aligned} \tag{C24}$$

Dividing by δM gives the derivative.

In our case, it is convenient to study:

$$\frac{1}{Z}\frac{d}{d\omega^2}Z = \sum_n \frac{1}{\omega^2+\omega_n^2}. \tag{C25}$$

This is progress. Our infinite product is now an infinite sum. The question is: how do we do the sum? The trick is to look for a periodic function which is well-behaved at infinity, but has poles at the integers. A suitable choice is

$$\frac{1}{e^{iz\beta}-1}. \tag{C26}$$

We can then replace any sum of the form, $\sum f(n)$, by a contour integral,

$$\frac{1}{2\pi}\int dz f(z)\frac{1}{e^{iz\beta}-1}. \tag{C27}$$

Here the contour is a line running just above the real z axis and back just below. The residues of the (infinite number of) poles just give back the original sum.

Now one can deform the contour, taking one line into the upper half plane, one into the lower, picking up the poles at $z=\pm i\omega$. This leaves us with:

$$\frac{dF}{d\omega^2} = \left(\frac{1}{e^{-\omega\beta}-1} - \frac{1}{e^{\omega\beta}-1}\right)\frac{1}{2\omega}. \tag{C28}$$

We could analyze this problem further, but let us jump instead to free field theory. Then

$$Z[\beta] = \int_{\phi(\beta)=\phi(0)} [d\phi] e^{-\int d^4x[(\partial_\mu\phi)^2+m^2\phi^2]}. \tag{C29}$$

In a finite box, with periodic boundary conditions, we can expand:

$$\phi(\vec{x},t) = \sum_{\vec{k},m} e^{i\vec{k}_n\cdot\vec{x}+i\omega_m t}\phi_{\vec{k},m} \tag{C30}$$

where $\omega_m = 2\pi m T$.

In this form, we have that

$$Z[\beta] = \det(-\partial^2+m^2)^{-1/2}. \tag{C31}$$

Again, this is somewhat awkward to work with. It is easier to differentiate:

$$\frac{1}{Z}\frac{\partial Z}{\partial m^2} = \frac{1}{Z}\int [d\phi] e^{-\int d^4x\mathcal{L}_E}\int d^4z\frac{1}{2}\phi^2(z). \tag{C32}$$

This is just the propagator, with periodic boundary conditions in the time direction:

$$\int d^4z \langle \phi(z)\phi(z)\rangle = \beta V \langle \phi(0)\phi(0)\rangle. \tag{C33}$$

The propagator is given by:

$$\langle \phi(0)\phi(0)\rangle = \sum_m \sum_k \frac{1}{\omega_m^2 + \vec{k}^2 + m^2}. \tag{C34}$$

We can convert this into a more recognizable form by means of the same trick. The propagator is given by the expression below:

$$\langle \phi(0)\phi(0)\rangle = \int \frac{d^3k}{(2\pi^2)^3} \frac{1}{2\pi} \int \frac{dz}{e^{iz\beta} - 1} \frac{1}{(2\pi nT)^2 + \vec{k}^2 + m^2}. \tag{C35}$$

Now deform the contour as before, picking up the poles at $\pm i\sqrt{\vec{k}^2 + m^2}$. Both poles make the same contribution, yielding:

$$\begin{aligned} &\frac{1}{2\sqrt{\vec{k}^2 + m^2}} \left[\frac{1}{e^{-\beta\sqrt{\vec{k}^2+m^2}} - 1} - \frac{1}{e^{\beta\sqrt{\vec{k}^2+m^2}} - 1} \right] \\ &\quad = \frac{1}{2\sqrt{\vec{k}^2 + m^2}} \left[1 + \frac{2}{e^{\beta\sqrt{\vec{k}^2+m^2}} - 1} \right]. \end{aligned} \tag{C36}$$

Note the appearance of the Bose–Einstein factors here. Note also the first term has the structure of the zero temperature expression for the energy; the second is the finite temperature expression. This is what we find differentiating:

$$\beta F = V \int \frac{d^3k}{(2\pi)^3} \left[\frac{1}{2} E_k + \beta^{-1} \ln(1 - e^{-\beta E_k}) \right]. \tag{C37}$$

Note the connection with the result for the single oscillator. So far our discussion has been for free field theory, but we can extend it immediately to interacting theories, developing a perturbation order by order in the couplings, just as at zero temperature.

C.3 QCD at high temperature

Two particularly important cases are QCD and the weak-interaction theory. At low energies, QCD is a complicated theory. But at high temperatures, things drastically simplify. In perturbation theory, if we are studying the free energy, for example, we are instructed to study a Euclidean problem with discrete energies which are multiples of T. So, provided that we do not encounter infrared problems, the free energy should be a power series in $g^2(T)$, calculable in perturbation theory.

One can argue that there is actually a phase transition between a confined phase and a deconfined phase. To find an order parameter for this transition, we start by considering a Wilson line, running from imaginary time $t = 0$ to $t = \beta$,

$$U_T(\vec{x}) = P e^{i \int_0^\beta A^0(\vec{x},t)dt}. \tag{C38}$$

Because of the periodic boundary conditions, this is gauge-invariant. The correlation of two such operators is related to the potential of two static quarks:

$$P(R) = \langle U_T(\vec{R}) U_T(0)\rangle = C\exp(-\beta V(R)). \tag{C39}$$

In a confining phase, with a linear potential between the quarks, $P(R)$ vanishes exponentially with R. In a Coulomb phase (nearly free quarks), it will tend to a constant. At very high temperatures, we would expect that we could compute P in a power series in $g^2(T)$, and that we will find the free quark behavior. Numerical studies show that there is a phase transition at a particular temperature between confined and unconfined phases. The order of the transition depends on the group.

Finite temperature perturbation theory suffers from infrared divergences, even at very high temperature. The problem is the zero-frequency modes in the sum over frequencies. If we simply set all of the frequencies to zero, we have the Feynman diagrams of a three-dimensional field theory. At four loops the divergence is logarithmic. At higher loops, it is power law.

We can understand this directly in the path integral. Consider a massless scalar field. The exponent in the path integral is:

$$\int_0^\beta dt d^3x (\partial_\mu \phi)^2. \tag{C40}$$

For small β, assuming it makes sense to treat fields as constant in β, the path integral thus becomes

$$\int [d\phi(\vec{x})]\, e^{-\beta H}, \tag{C41}$$

which is the classical partition function for the three-dimensional system.

Thought of in this way, there is a natural guess for how the infrared divergences are cut off. A three-dimensional gauge theory has a dimensionful coupling λ^2. One might expect that such a theory has a mass gap proportional to λ^2 (in three dimensions, the gauge coupling has dimensions of $\sqrt{M}$). In the present case the coupling is $\lambda = g^2 T$. This scale then would cut off the infrared divergence. This suggests that the theory at finite temperature makes sense, but does not help a great deal with computations. The problem is that in four loops, we obtain a contribution $g^8 \ln(g^2)$, but at higher orders we obtain a power series in g^2/g^2, i.e. we can at best compute the leading logarithmic term at four loops. It is possible to study some of these issues numerically in lattice gauge theory, which provides some support for this picture.

Instanton effects at high temperature

In QCD at zero temperature, we saw that instanton calculations were plagued by infrared divergences. At high temperatures, this is not the case. The scale invariance of the zero-energy theory is lost, and the instanton solution has a definite scale, of order the temperature. As a result, instanton effects behave as $\exp(-8\pi^2/g^2(T))$, and are calculable. Thus it is possible to compute the θ dependence systematically. This is particularly relevant to the understanding of the axion in the early universe.

C.4 Weak interactions at high temperature

The weak interactions exhibit different phenomena at high temperatures. Most strikingly, there is a transition between a phase in which the gauge bosons are massive and one in which they are massless. This transition can be uncovered in perturbation theory. By analogy with the phase transition in the Landau–Ginzburg model of superconductivity, one might expect that the value of $\langle \Phi \rangle$ will change as the temperature increases. To

determine the value of Φ, one must compute the free energy as a function of Φ. The leading temperature-dependent corrections are obtained by simply noting that the masses of the various fields in the theory (the W and Z bosons and the Higgs field, in particular) depend on Φ. So the contributions of each species to the free energy are Φ-dependent:

$$\mathcal{F}(\Phi)V_T(\Phi) = \pm \sum_i \int \frac{d^3p}{2\pi^3} \ln\left(1 \mp e^{-\beta\sqrt{p^2+m_i^2(\Phi)}}\right), \tag{C42}$$

where $\beta = 1/T$, T is the temperature, the sum is over all particle species (physical helicity states), and the plus sign is for bosons, the minus for fermions. In the Standard Model, for temperature $T \sim 10^2$ GeV, one can treat all the quarks as massless, except for the top quark. The effective potential (C42) then depends on the top quark mass, m_t, the vector boson masses, M_z and m_w, and on the Higgs mass, m_H. Performing the integral in the equation yields

$$V(\Phi, T) = D\left(T^2 - T_0^2\right)\Phi^2 - ET\Phi^3 + \frac{\lambda}{4}\Phi^4 + \cdots . \tag{C43}$$

The parameters T_0, D and E are given in terms of the gauge boson masses and the gauge couplings. For the moment, though, it is useful to note certain features of this expression. E turns out to be a rather small, dimensionless number, of order 10^{-2}. If we ignore the ϕ^3 term, we have a second-order transition, at temperature T_0, between a phase with $\phi \neq 0$ and a phase with $\phi = 0$. Because the W and Z masses are proportional to ϕ, this is a transition between a state with massive and massless gauge bosons.

Because of the ϕ^3 term in the potential, the phase transition is potentially at least weakly first order. A second, distinct, minimum appears at a critical temperature. A first-order transition is not, in general, an adiabatic process. As we lower the temperature to the transition temperature, the transition proceeds by the formation of bubbles; inside the bubble the system is in the true equilibrium state (the state which minimizes the free energy) while outside it tends to the original state. These bubbles form through thermal fluctuations at different points in the system, and grow until they collide, completing the phase transition. The moving bubble walls are regions where the Higgs fields are changing, and all of Sakharov's conditions are satisfied.

C.5 Electroweak baryon number violation

We have seen that, at low temperatures, violations of baryon and lepton number are extremely small. This is not the case at high temperature, where baryon number violation is a rapid process, which can come to thermal equilibrium. This has at least two possible implications. First, it is conceivable that these *sphaleron* processes can themselves be responsible for generating a baryon asymmetry. This is called electroweak baryogenesis. Second, sphaleron processes can process an existing lepton number, producing a net lepton and baryon number. This is the process called leptogenesis. In this section, we summarize the main arguments that the electroweak interactions violate baryon number at high temperature.

Recall that, classically, the ground states are field configurations for which the energy vanishes. The trivial solution of this condition is $\vec{A} = 0$, where $\vec{A}$ is the vector potential. More generally, one can consider $\vec{A}$ which is a "pure gauge,"

$$\vec{A} = \frac{1}{i} g^{-1} \vec{\nabla} g, \tag{C44}$$

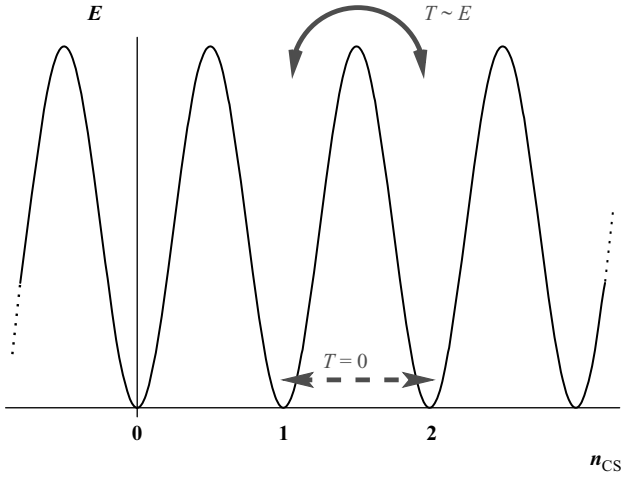

Fig. C1. Schematic Yang–Mills vacuum structure. At zero temperature, the instanton transitions between vacua with different Chern–Simons numbers are suppressed. At finite temperature, these transitions can proceed via sphalerons.

where g is a gauge transformation matrix. In an Abelian ($U(1)$) gauge theory, fixing the gauge eliminates all but the trivial solution, $\vec{A} = 0$.[1] This is not the case for non-Abelian gauge theories. There is a class of gauge transformations, labeled by a discrete index n, which do not tend to unity as $|\vec{x}| \to \infty$, which must be considered to be distinct states. These have the form:

$$g_n(\vec{x}) = e^{inf(\vec{x})\hat{x}\cdot\tau/2}, \tag{C45}$$

where $f(x) \to 2\pi$ as $\vec{x} \to \infty$, and $f(\vec{x}) \to 0$ as $\vec{x} \to 0$.

So the ground states of the gauge theory are labeled by an integer n. Now if we evaluate the integral of the current K^0, we obtain a quantity known as the Chern–Simons number:

$$n_{\rm CS} = \frac{1}{16\pi^2}\int d^3x K^0 = \frac{2/3}{16\pi^2}\int d^3x \epsilon_{ijk}{\rm Tr}(g^{-1}\partial_i g g^{-1}\partial_j g g^{-1}\partial_k g). \tag{C46}$$

For $g = g_n$, $n_{\rm CS} = n$. The reader can also check that for $g' = g_n(x)h(x)$, where h is a gauge transformation which tends to unity at infinity (a so-called "small gauge transformation"), this quantity is unchanged. The "Chern–Simons number," $n_{\rm CS}$, is topological in this sense (for $\vec{A}$s which are not "pure gauge," $n_{\rm CS}$ is in no sense quantized).

Schematically, we can thus think of the vacuum structure of a Yang–Mills theory as indicated in Fig. C1. We have, at weak coupling, an infinite set of states, labeled by integers, and separated by barriers from one another. In tunneling processes which change the Chern–Simons number, because of the anomaly, the baryon and lepton numbers will change. The exponential suppression found in the instanton calculation is typical of tunneling processes, and in fact the instanton calculation which leads to the result for the amplitude is nothing but a field-theoretic WKB calculation.

[1] More precisely, this is true in axial gauge. In the gauge $A_0 = 0$, it is necessary to sum over all time-independent transformations to construct a state which obeys Gauss's law.

One can determine the height of the barrier separating configurations of different $n_{\rm CS}$ by looking for the field configuration which corresponds to sitting on top of the barrier. This is a solution of the static equations of motion with finite energy. It is known as a "sphaleron." When one studies the small fluctuations about this solution, one finds that there is a single negative mode, corresponding to the possibility of rolling down hill into one or the other well. The sphaleron energy is of order

$$E_{\rm sp} = \frac{c}{g^2} M_W . \tag{C47}$$

This can be seen by scaling arguments on the classical equations; determining the coefficient c requires a more detailed analysis. The rate for thermal fluctuations to cross the barrier per unit time per unit volume should be of order the Boltzmann factor for this configuration, multiplied by a suitable prefactor,

$$\Gamma_{\rm sp} = T^4 e^{-E_{\rm sp}/T} . \tag{C48}$$

Note that the rate becomes large as the temperature approaches the W boson mass. The W boson mass itself goes to zero as one approaches the electroweak phase transition. At this point, the computation of the transition rate is a difficult problem – there is no small parameter – but general scaling arguments show that the transition rate is of the form:[2]

$$\Gamma_{\rm bv} = \alpha_w^4 T^4 . \tag{C49}$$

Suggested reading

The path integral is well treated in most modern field theory textbooks. Peskin and Schroder (1995) provide a concise introduction. High-temperature field theory is developed in a number of textbooks, such as that of Kapusta (1989).

Exercises

(1) Go through the calculation of the free energy of a free scalar field carefully, being careful about factors of 2 and π.
(2) Compute the constants appearing in Eq. (C43). Plot the free energy, and show that the transition is weakly first order.
(3) Show, by power counting, that infrared divergences first appear in the free energy of a gauge theory at three loops. To do this, you can look at the zero-frequency terms in the sums over frequency. Show that the divergences become more severe at higher orders.

[2] More detailed considerations alter slightly the parametric form of the rate.

Appendix D

The beta function in supersymmetric Yang–Mills theory

We have seen that holomorphy is a powerful tool to understand the dynamics of supersymmetric field theories. But one can easily run into puzzles and paradoxes. One source of confusion is the holomorphy of the gauge coupling. At tree level, the gauge coupling arises from a term in the action of the form:

$$\int d^2\theta S W_\alpha^2 \tag{D1}$$

where $S = -(1/4g^2) + ia$. This action, in perturbation theory, has a symmetry

$$S \to S + i\alpha. \tag{D2}$$

This is just an axion shift symmetry. Combined with holomorphy, this greatly restricts the form of the effective action. The only allowed terms are:

$$\mathcal{L}_{\text{eff}} = \int d^2\theta (S + \text{constant}) W_\alpha^2. \tag{D3}$$

The constant term corresponds to a one-loop correction. But higher-loop corrections are forbidden.

On the other hand, it is well known that there are two-loop corrections to the beta function in supersymmetric Yang–Mills theories (higher-loop corrections have also been computed). Does this represent an inconsistency? This puzzle can be stated – and has been stated – in other ways. For example, the axial anomaly is in a supermultiplet with the conformal anomaly – the anomaly in the trace of the stress tensor. One usually says that the axial anomaly is not renormalized, but the trace anomaly is proportional to the beta function.

The resolution to this puzzle was provided by Shifman and Vainshtein. It is most easily described in a $U(1)$ gauge theory, with some charged superfields, say $\phi^\pm$. Without masses for these fields, the one-particle irreducible effective action has infrared singularities. In addition, we need to regulate ultraviolet divergences. We can regulate the second type of divergence by introducing Pauli–Villars regulator fields, while the infrared divergence can be regulated by including a mass for $\phi^\pm$. The $\phi^\pm$ and regulator mass terms are holomorphic:

$$\int d^2\theta M \Phi_{\text{pv}}^+ \Phi_{\text{pv}}^-. \tag{D4}$$

The gauge coupling term in the effective action must be a holomorphic function, now, of S and M. But the effective action also includes wave function renormalizations for the various regulator fields:

$$\int d^4\theta Z^{-1}(\Phi^{+\dagger}\Phi^{+} + \Phi^{-\dagger}\Phi^{-}). \tag{D5}$$

The wave function factors, Z, are *not* holomorphic functions of the parameters.[1] The physical cutoff is then $Z_{\rm pv}M$, and the physical infrared scale is $Z_\phi m$. So to determine the coupling constant renormalization in terms of this scale, we need to compute the Zs as well. One needs to be a bit careful in this computation. If one works in a non-supersymmetric gauge, such as Wess–Zumino gauge, one needs to actually compute the mass renormalization; the wave function renormalizations will be different for the different component fields.

Starting, then, with our holomorphic expression

$$\frac{8\pi^2}{g^2(m)} = \frac{8\pi^2}{g^2(\Lambda)} + b_0 \ln(m/M), \tag{D6}$$

we have, in terms of the physical masses:

$$\frac{8\pi^2}{g^2(m)} = \frac{8\pi^2}{g^2(\Lambda)} + b_0(\ln(m/M) - \ln(Z(m)/Z(M))). \tag{D7}$$

To form the beta function, we need to take

$$\beta(g) = \frac{-g^3}{16\pi^2}\frac{\partial}{\partial \ln(m)} g^{-2}(m) = -\frac{b_0 g^3}{16\pi^2} + \frac{b_0 g^3}{16\pi^2}\gamma \tag{D8}$$

where

$$\gamma = \frac{d}{d\ln(M)}\ln(Z) = -\frac{4g^2}{16\pi^2}. \tag{D9}$$

So there are two- and higher-loop corrections to the beta function. Plugging in, one obtains to two-loop order, for the $U(1)$ theory:

$$-\beta(g) = \frac{g^3}{16\pi^2} + \frac{4g^5}{(16\pi^2)^2}. \tag{D10}$$

This is, in fact, the correct result.

This analysis makes clear why the holomorphic analysis is correct but subtle. For non-Abelian theories, it is not quite so straightforward to introduce a holomorphic regulator. One can arrive at the required modification by a variety of arguments. In many ways, the most convincing and straightforward comes from examination of instanton amplitudes. One can also simply make an educated guess by examining the results of two-loop computations. The required relation is:

$$\frac{8\pi^2}{g^2(m)} = \frac{8\pi^2}{g^2(\Lambda)} + b_0(\ln(m/M) - \ln(Z(m)/Z(M)) + C_A \ln(g^2). \tag{D11}$$

[1] There is much dispute in the literature about whether to write the action with Z or Z^{-1}. I have chosen Z^{-1}, following the convention of most field theory texts, in which propagators have a factor of Z in the numerator. The reader is free to follow his or her taste.

Differentiating, as before, one obtains the expression for the beta function:

$$\beta(g) = \frac{g^3}{8\pi^2} \frac{3C_A - \sum T_F^i (1 - \gamma^i)}{1 - (C_A g^2/8\pi)^2}. \tag{D12}$$

Exercise

(1) By examining 't Hooft's computation of the instanton determinant, argue that the appropriate generalization of the Shifman–Vainshtein formula is that of Eq. (D11). Derive the exact expression for the beta function of Eq. (D12). Verify, by comparison with published results, that this correctly reproduces the two-loop beta function of a supersymmetric gauge theory.

References

Affleck, I. (1980). *Phys. Lett.*, **B92**, 149.

Affleck, I., Dine, M. and Seiberg, N. (1984). Dynamical supersymmetry breaking in supersymmetric QCD. *Nucl. Phys. B*, **241**, 493.

Aharony, O., Gubser, S. S., Maldacena, J. M., Ooguri, H. and Oz, Y. (2000). Large N field theories, string theory and gravity. *Phys. Rept.*, **323**, 183 [arXiv:hep-th/9905111].

Albrecht, A. and Steinhardt, P. J. (1982). Cosmology for grand unified theories with radiatively induced symmetry breaking. *Phys. Rev. Lett.*, **48**, 1220.

Appelquist, T., Chodos, A. and Freund, P. G. O. (1985). *Modern Kaluza–Klein Theories*. Menlo Park: Benjamin/Cummings.

Arkani-Hamed, N., Dimopoulos, S. and Dvali, G. R. (1999). Phenomenology, astrophysics and cosmology of theories with sub-millimeter dimensions and TeV scale quantum gravity. *Phys. Rev. D*, **59**, 086004 [arXiv:hep-ph/9807344].

Ashok, S. and Douglas, M. R. (2004). Counting flux vacua. *JHEP*, **0401**, 060 [arXiv:hep-th/0307049].

Atick, J. J., Dixon, L. J. and Sen, A. (1987). String calculation of Fayet–Iliopoulos D terms in arbitrary supersymmetric compactifications. *Nucl. Phys. B*, **292**, 109.

Bagger, J. A., Moroi, T. and Poppitz, E. (2000). Anomaly mediation in supergravity theories. *JHEP*, **0004**, 009 [arXiv:hep-th/9911029].

Bailin, D. and Love, A. (1993). *Introduction to Gauge Field Theory*. London: Institute of Physics.

Banks, T., Dine, M. and Graesser, M. (2003). Supersymmetry, axions and cosmology. *Phys. Rev. D*, **68**, 075011 [arXiv:hep-ph/0210256].

Banks, T. and Dixon, L. J. (1988). Constraints on string vacua with space-time supersymmetry. *Nucl. Phys. B*, **307**, 93.

Banks, T., Fischler, W., Shenker, S. H. and Susskind, L. (1997). M theory as a matrix model: A conjecture. *Phys. Rev. D*, **55**, 5112 [arXiv:hep-th/9610043].

Banks, T., Kaplan, D. B. and Nelson, A. E. (1994). Cosmological implications of dynamical supersymmetry breaking. *Phys. Rev. D*, **49**, 779 [hep-ph/9308292].

Berkooz, M., Dine, M. and Volansky, T. (2004). Hybrid inflation and the moduli problem. *Phys. Rev. D*, **71**, 103 502 [arXiv:hep-ph/0409226].

Bigatti, D. and Susskind, L. (1997). Review of matrix theory. In *Strings, Branes and Dualities*, Proceedings, ed. L. Balieu, Philippe Di Francesco, M. Douglas *et al*. Amsterdam: Kluwer [arXiv:hep-th/9712072].

Bousso, R. and Polchinski, J. (2000). Quantization of four-form fluxes and dynamical neutralization of the cosmological constant. *JHEP*, **0006**, 006 [arXiv:hep-th/0004134].
Brown, L. S., Carlitz, R. D., Creamer, D. B. and Lee, C. K. (1978). Propagation functions in pseudoparticle fields. *Phys. Rev. D*, **17**, 1583.
Buchmuller, W., Di Bari, P. and Plumacher, M. (2005). Leptogenesis for pedestrians. *Annals Phys.*, **315**, 305 [arXiv:hep-ph/0401240].
Carena, M., Quiros, M., Seco, M. and Wagner, C. E. M. (2003). Improved results in supersymmetric electroweak baryogenesis. *Nucl. Phys. B*, **650**, 24 [arXiv:hep-ph/0208043].
Carroll, S. (2004). *Spacetime and Geometry: An Introduction to General Relativity*. San Francisco: Addison-Wesley.
Casher, A., Kogut, J. B. and Susskind, L. (1974). Vacuum polarization and the absence of free quarks. *Phys. Rev. D*, **10**, 732.
Cheng, T. and Li, L. (1984). *Gauge Theory of Elementary Particle Physics*. Oxford: Clarendon Press.
Cheng, T.-P. and Li, L.-F. (1984). *Gauge Theory of Elementary Particle Physics*. Oxford: Clarendon Press.
Chivukula, R. S. Technicolor and compositeness, arXiv:hep-ph/0011264.
Cohen, A. G., Kaplan, D. B. and Nelson, A. E. (1993). Progress in electroweak baryogenesis. *Ann. Rev. Nucl. Part. Sci.*, **43**, 27 [arXiv: hep-ph/9302210].
Coleman, S. (1985). The uses of instantons, in *Aspects of Symmetry*. Cambridge: Cambridge University Press.
Coleman, S. R. (1981). The magnetic monopole fifty years later, in Zichichi, A. (1983). *The Unity of the fundamental interactions*. New York: Plenum Press.
Cottingham, W. N. and Greenwood, D. A. (1998). *An Introduction to the Standard Model of Particle Physics*. Cambridge: Cambridge University Press.
Cremmer, E., Julia, B., Scherk, J. *et al*. (1979). Spontaneous symmetry breaking and Higgs effect in supergravity without cosmological constant. *Nuclear Physics B*, **147**, 105.
Creutz, M. (1983). *Quarks, Gluons and Lattices*. Cambridge: Cambridge University Press.
Crewther, R. J., Di Vecchia, P., Veneziano, G. and Witten, E. (1979). *Phys. Lett.*, **88B**, 123.
Davoudiasl, H., Hewett, J. L. and Rizzo, T. G. (2000). Phenomenology of the Randall–Sundrum gauge hierarchy model. *Phys. Rev. Lett.*, **84**, 2080 [arXiv:hep-ph/9909255].
de Carlos, B., Casas, J. A., Quevedo, F. and Roulet, E. (1993). Model independent properties and cosmological implications of the dilaton and moduli sectors of 4-d strings. *Phys. Lett. B*, **318**, 447 [arXiv:hep-ph/9308325].
D'Hoker, E. and Freedman, D. Z. (2002). Supersymmetric gauge theories and the AdS/CFT correspondence [arXiv:hep-th/0201253].
Dimopoulos, S. and Georgi, H. (1981). Softly broken supersymmetry and SU(5). *Nucl. Phys. B*, **193**, 150.
Dine, M. and Kusenko, A. (2003). The origin of the matter-antimatter asymmetry. *Rev. Mod. Phys.*, **76**, 1 [arXiv:hep-ph/0303065].
Dine, M. and Seiberg, N. (1986). Nonrenormalization theorems in superstring theory. *Phys. Rev. Lett.*, **57**, 2625.
Dine, M., Seiberg, N., Wen, X. G. and Witten, E. (1986). Nonperturbative effects on the string world sheet. *Nucl. Phys. B*, **278**, 769.
Dine, M., Ichinose, I. and Seiberg, N. (1987a). F terms and D terms in string theory. *Nucl. Phys. B*, **293**, 253.

Dine, M., Seiberg, N., Wen, X. G. and Witten, E. (1987b). Nonperturbative effects on the string world sheet. 2. *Nucl. Phys. B*, **289**, 319.
Dine, M., Seiberg, N. and Witten, E. (1987c). Fayet–Iliopoulos terms in string theory. *Nucl. Phys. B*, **289**, 589.
Distler, J. and Greene, B. R. (1988). Some exact results on the superpotential from Calabi–Yau compactifications. *Nucl. Phys. B*, **309**, 295.
Dixon, L. J., Harvey, J. A., Vafa, C. and Witten, E. (1986). Strings on orbifolds. 2. *Nucl. Phys. B*, **274**, 285.
Dodelson, S. (2004). *Modern Cosmology*. Burlington: Academic Press.
Donoghue, J. F., Golowich, E. and Holstein, B. R. (1992). *Dynamics of the Standard Model*. Cambridge: Cambridge University Press.
Eidelman, S. *et al.* [Particle Data Group] (2004). Review of particle physics. *Phys. Lett. B*, **592**, 1 (This article – and others – can be found on the Particle Data Group website).
Enqvist, K. and Mazumdar, A. (2003). Cosmological consequences of MSSM flat directions. *Phys. Rept.*, **380**, 99 [arXiv:hep-ph/0209244].
Faraggi, A. E. (1999). Toward the classification of the realistic free fermionic models. *Int. J. Mod. Phys. A*, **14**, 1663 [arXiv:hep-th/9708112].
Feng., J. L., March-Russell, J., Sethi, S. and Wilczek, F. (2001). Saltatory relaxation of the cosmological constant. *Nucl. Phys. B*, **602**, 307 [arXiv:hep-th/0005276].
Fradkin, E. and Shenker, S. (1979). Phase diagrams of lattice gauge theories with Higgs fields. *Phys. Rev. D*, **19**, 3602.
Gates, S. J., Grisaru, M. R. and Siegel, W. (1983). *Superspace: or One Thousand and One Lessons in Supersymmetry*. San Francisco: Benjamin/Cummings.
Gepner, D. (1987). Exactly solvable string compactifications on manifolds of SU(N) holonomy. *Phys. Lett. B*, **199**, 380.
Giudice, G. F. and Rattazzi, R. (1999). Theories with gauge-mediated supersymmetry breaking. *Phys. Rept.*, **322**, 419 [arXiv:hep-ph/9801271].
Glashow, S. L., Iliopoulos, J. and Maiani, L. (1970). Weak interactions with lepton–hadron symmetry. *Phys. Rev. D*, **2**, 1285.
Green, M. B., Schwarz, J. H. and Witten, E. (1987). *Superstring Theory,* Cambridge: Cambridge University Press.
Greene, B. R., Kirklin, K. H., Miron, P. J. and Ross, G. G. (1987) A three generation superstring model. 2. Symmetry breaking and the low-energy theory, *Nucl. Phys. B*, **292**, 606.
Gross, D. J., Harvey, J. A., Martinec, E. J. and Rohm, R. (1985). Heterotic string theory. 1. The free heterotic string. *Nucl. Phys. B*, **256**, 253.
Gross, D. J., Harvey, J. A., Martinec, E. J. and Rohm, R. (1986). Heterotic string theory. 2. The interacting heterotic string. *Nucl. Phys. B*, **267**, 75.
Gross, D. J. and Wilczek, F. (1973). Ultraviolet behavior of non-abelian gauge theories. *Phys. Rev. Lett.*, **30**, 1343.
Guth, A. H. (1981). The inflationary universe: a possible solution to the horizon and flatness problems. *Phys. Rev. D*, **23**, 347.
Hartle, J. B. (2003). *Gravity, an Introduction to Einstein's General Relativity*. San Francisco: Addison-Wesley.
Harvey, J. (1996). Magnetic monopoles, duality, and supersymmetry, arXiv:hep-th/9603086.
Intriligator, K. A. and Seiberg, N. (1996). Lectures on supersymmetric gauge theories and electric-magnetic duality. *Nucl. Phys. Proc. Suppl.*, **45BC**, 1 [arXiv:hep-th/9509066].

Intriligator, K. and Thomas, S. (1996). Dynamical supersymmetry breaking on quantum moduli spaces. *Nucl. Phys.*, **B473**, 121, arXiv:hep-th/9603158.
Jackson, J. D. (1999). *Classical Electrodynamics*. Hoboken: Wiley.
Johnson, C. (2003). *D-Branes*. Cambridge: Cambridge University Press.
Kachru, S., Kallosh, R., Linde, A. and Trivedi, S. (2003). De Sitter vacua in string theory. *Phys. Rev. D*, **68**, 046005 [arXiv:hep-th/0301240].
Kapusta, J. I. (1989). *Finite-Temperature Field Theory*. Cambridge: Cambridge University Press.
Kolb, E. W. and Turner, M. S. (1990). *The Early Universe*. Redwood City: Addison-Wesley.
Linde, A. D. (1982). A new inflationary universe scenario: a possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problems. *Phys. Lett.*, **108B**.
Linde, A. (1990). *Particle Physics and Inflationary Cosmology*. Reading: Harwood Academic.
Linde, A. D. (1994). Hybrid inflation. *Phys. Rev. D*, **49**, 748 [arXiv:astro-ph/9307002].
Lykken, J. D. (1996). Introduction to supersymmetry, arXiv:hep-th/9612114.
Maldacena, J. M. (1997). The large N limit of superconformal field theories and supergravity. *Adv. Theor. Math. Phys.*, **2**, 231 [arXiv:hep-th/9711200].
Manohar, A. V. and Wise, M. B. (2000). *Heavy Quark Physics*. Cambridge: Cambridge University Press.
Martin, S. P. and Vaughn, M. T. (1994). Two loop renormalization group equations for soft supersymmetry breaking couplings. *Phys. Rev. D*, **50**, 2282 [arXiv:hep-ph/9311340].
Masiero, A. and Silvestrini, L. (1997). Two lectures on FCNC and CP violation in supersymmetry, arXiv:hep-ph/9711401.
Mohapatra, R. N. (2003). *Unification and Supersymmetry: The Frontiers of Quark – Lepton Physics*. Berlin: Springer-Verlag.
Nilles, H. P. (1984). Supersymmetry and supergravity. *Phys. Rept*, **110**, 1.
Olive, D. I. and Witten, E. (1978). Supersymmetry algebras that include topological charges. *Phys. Lett. B*, **78**, 97.
Pais, A. (1986). *Inward Bound*. Oxford: Clarendon Press.
Peet, A. (2000). TASI lectures on black holes in string theory [arXiv:hep-th/0008241].
Peskin, M. E. Introduction to string and superstring theory. 2, SLAC-PUB-4251.
Peskin, M. E. (1985). SLAC-PUB-3821 *Lectures Given at Theoretical Advanced Study Institute on Elementary Particle Physics, New Haven, CT, June 9–July 5, 1985*.
Peskin, M. E. (1987). In *From the Planck Scale to the Weak Scale: Towards a Theory of the Universe*, ed. H. E. Haber. Singapore: World Scientific.
Peskin, M. E. (1990). *Theory of Precision Electroweak Measurements*. Lectures given at 17th SLAC Summer Inst.: Physics at the 100-GeV Mass Scale, Stanford, CA.
Peskin, M. E. (1997). In *Fields, Strings and Duality: TASI 96*, ed. C. Efthimiouo and B. Greene. Singapore: World Scientific.
Peskin, M. E. and Schroeder, D. V. (1995). *An Introduction to Quantum Field Theory*. Menlo Park: Addison Wesley.
Peskin, M. E. and Takeuchi, T. (1990). A new constraint on a strongly interacting Higgs sector. *Phys. Rev. Lett.*, **65**, 964.
Pokorski, S. (2000). *Gauge Field Theories*. Cambridge: Cambridge University Press.
Polchinski, J. (1998). *String Theory*. Cambridge: Cambridge University Press.
Politzer, H. D. (1973). Reliable perturbative results for strong interactions? *Phys. Rev. Lett.*, **30**, 1346.
Ramond, P. (1999). *Journeys Beyond the Standard Model*. New York: Perseus Books.

Randall, L., Soljacic, M. and Guth, A. H. (1996). Supernatural inflation: inflation from supersymmetry with no (very) small parameters. *Nucl. Phys. B*, **472**, 377 [arXiv:hep-ph/9512439].
Randall, L. and Sundrum, R. (1999). A large mass hierarchy from a small extra dimension. *Phys. Rev. Lett.*, **83**, 3370 [arXiv:hep-ph/9905221].
Rohm, R. (1984). Spontaneous supersymmetry breaking in supersymmetric string theories. *Nucl. Phys. B*, **237**, 553.
Ross, G. G. (1984). *Grand Unified Theories*. Boulder: Westview Press.
Salam, A. and Ward, J. C. (1964). Electromagnetic and weak interactions. *Phys. Lett.*, **13**, 168.
Sannan, S. (1986). Gravity as the limit of the type II superstring theory. *Phys. Rev. D*, **34**, 1749.
Seiberg, N. (1993). Naturalness versus supersymmetric nonrenormalization theorems. *Phys. Lett. B*, **318**, 469 [arXiv:hep-ph/9309335].
Seiberg, N. (1994a). The power of holomorphy: exact results in 4-D SUSY field theories. In *International Symposium on Particles, Strings and Cosmology (PASCOS94)*, ed. K. C. Wali. Singapore: World Scientific [arXiv:hep-th/9408013].
Seiberg, N. (1994b). Exact results on the space of vacua of four-dimensional SUSY gauge theories. *Phys. Rev. D*, **49**, 6857 [arXiv:hep-th/9402044].
Seiberg, N. (1995a). The power of duality: exact results in 4D SUSY field theory. *Int. J. Mod. Phys. A*, **16**, 4365 [arXiv:hep-th/9506077].
Seiberg, N. (1995b). Electric–magnetic duality in supersymmetric nonAbelian gauge theories. *Nucl. Phys. B*, **435**, 129 [arXiv:hep-th/9411149].
Seiberg, N. (1997). Why is the matrix model correct? *Phys. Rev. Lett.*, **79**, 3577 [arXiv:hep-th/9710009].
Seiberg, N. and Witten, E. (1994). Electric–magnetic duality, monopole condensation, and confinement in $N = 2$ supersymmetric Yang–Mills theory. *Nucl. Phys. B*, **426**, 19 [Erratum-ibid. *B*, **430**, 485 (1994)] [arXiv:hep-th/9407087].
Seiden, A. (2005). *Particle Physics: A Comprehensive Introduction*. San Francisco: Addison-Wesley.
Shadmi, Y. and Shirman, Y. (2000). Dynamical supersymmetry breaking. *Rev. Mod. Phys.*, **72**, 25 [arXiv:hep-th/9907225].
Silverstein, E. and Witten, E. (1995). Criteria for conformal invariance of (0, 2) models. *Nucl. Phys. B*, **444**, 161 [arXiv:hep-th/9503212].
Susskind, L. (1977). Coarse grained quantum chromodynamics. In *Weak and Electromagnetic Interactions at High Energy*, ed. R. Balian and C. H. Llewelleyn Smith. Amsterdam: North-Holland.
Terning, J. (2003). Non-perturbative supersymmetry, arXiv:hep-th/0306119.
't Hooft, G. (1971). Renormalization of massless Yang–Mills fields. *Nucl. Phys. B*, **33**, 173.
't Hooft, G. (1976). *Phys.Rev. D***14**, 3432; erratum-*ibid*. (1978) *D***18**, 2199.
't Hooft, G. (1980). In *Recent Developments in Gauge Theories*, ed. G. 't Hooft *et al*. New York: Plenum Press.
Turner, M. S. (1990). Windows on the axion. *Phys. Rept.*, **197**, 67.
Vilenkin, A. (1995). Predictions from quantum cosmology. *Phys. Rev. Lett.*, **74**, 846 [arXiv:gr-qc/9406010].
Wald, R. M. (1984). *General Relativity*. Chicago: University of Chicago Press.
Weinberg, S. (1967). A model of leptons. *Phys. Rev. Lett.*, **19**, 1264.
Weinberg, S. (1972). *Gravitation and Cosmology, Principles and Applications of the General Theory of Relativity*. New York: John Wiley and Sons.

Weinberg, S. (1989). The cosmological constant problem. *Rev. Mod. Phys.*, **61**, 1.
Weinberg, S. (1995). *The Quantum Theory of Fields*. Cambridge: Cambridge University Press.
Weinberg, S. (2000). The cosmological constant problems [arXiv:astro-ph/0005265].
Wess, J. and Bagger, J. (1992). *Supersymmetry and Supergravity*, Princeton: Princeton University Press.
Wilson, K. G. (1974). Confinement of quarks. *Phys. Rev. D*, **10**, 2445.
Witten, E. (1981). Dynamical breaking of supersymmetry. *Nucl. Phys. B*, **188**, 513.
Witten, E. (1986). New issues in manifolds of SU(3) holonomy. *Nucl. Phys. B*, **268**, 79.
Witten, E. (1995). String theory dynamics in various dimensions. *Nucl. Phys. B*, **443**, 85 [e-Print Archive: hep-th/9503124].
Witten, E. (1998). Anti-de Sitter space and holography. *Adv. Theor. Math. Phys.*, **2**, 253 [arXiv:hep-th/9802150].
Yang, C. N. and Mills, R. L. (1954). Conservation of isotopic spin and isotopic gauge invariance. *Phys. Rev.*, **96**, 191.

Index

α' expansion, 429
Abelian Higgs model, 122
ADM energy, 254
AdS space, 464
AdS-CFT correspondence, 463, 464
Affleck, I., 198
Affleck–Dine baryogenesis, 287, 292, 293, 294, 295, 296, 299
Alvarez-Gaume, L., 308
anomalies, 75
anomalies, two dimensions, 81
anomaly, 77, 78, 79
anomaly by point splitting, 79
anomaly evaluation, Fujikawa, 83
anomaly matching conditions, 238
anomaly mediation, 213
anomaly, applications, 80
anomaly, Fujikawa evaluation, 80
Appelquist, T., 6
area law, 49
asymptotic freedom, 6, 35
Atiyah–Singer index theorem, 95
ATLAS, 475
axion, dark matter, 280
axion, finite temperature potential, 282
axion, stars, 281
axions, 103, 104, 105

background field gauge, 41
background field method, 40
Banks, T., 234, 277, 462, 479
Banks–Zaks fixed point, 234
baryogenesis, 270, 287
baryogenesis, gut, 287
baryogenesis, Sakharov conditions, 287
baryon asymmetry, 243
baryon number conservation, 63
Bekenstein, J., 255
beta function, 36, 39
beta function, non-linear sigma-model, 384, 385
beta function, sigma-model, 383
beta function, supersymmetric theories, 501
Betti numbers, 404
Bianchi identities, 261
Bianchi identity, 250
Bjorken, J., 4, 57
black holes, 255
Boltzmann equation, 284
bosonic closed string, spectrum, 319
bosonic open string, spectrum, 318
bosonic string, 313
bosonization, 387
BPS relations, 225

Calabi–Yau manifolds, spectra, 409
Calabi–Yau manifolds, three-form, 415
Calabi–Yau spaces, 401
Calabi–Yau spaces: Standard Model gauge group, 424
Casimir energy, 375
charge quantization, 113
Chern class, 405, 409
Chern–Simons number, 498
chiral basis, 483, 484
chiral superfields, 142
Christoffel connection, 247
CKM matrix, 25, 27, 33
CMS, 475
Coleman, S., 6
Coleman–Mandula theorem, 139
collective coordinate, 122
collective coordinates, 89, 94, 121, 125
confinement, 18, 44, 45
conformal field theories, 233
conformal field theory, 323, 325
conformal field theory, commutators, 326
conformal field theory, stress tensor, 321
conformal gauge, 320
continuous symmetries in string theory, 416
contravariant vector, 246
cosmic microwave background radiation, CMBR, 266
cosmological constant, 72
cosmological principle, 259
coupling constant unification, 186
covariant derivative, gauge theory, 10
covariant vector, 245
covariantly constant spinor, 407

CP^N model, 84, 85, 86, 88, 89
curvature tensor, 249

D-brane, actions, 447
D-brane charges, 446
D-branes, 443, 444, 445
D-terms, criteria for vanishing, 195
dark energy, 72, 243, 262, 269, 300
dark matter, 72, 243, 262, 267, 269
de Sitter space, 262
Dedekind η-functions, 337
deep inelastic scattering, 57
determinant, derivative of, 247
differential forms, 401, 403
dilaton, as string coupling constant, 369
dimension-five operators and proton decay, 188
dimension-six operators, 69
dimensional regularization, 40
dimensional transmutation, 44, 85
Dine, M., 198
Dirac, P. A. M., 119, 122
Dirac quantization condition, 120
Dirac strings, 719
Dirichlet boundary conditions, 316, 444
discrete symmetries, 168
discrete symmetries in string theory, 416, 422
Dolbeault operators, 405
Douglas, M., 462
dsb, criteria, 209
duality, IIA–11-dimensional supergravity, 452, 453, 454
duality, perturbative, 442
duality, strong, weak coupling, 443, 451
dynamical supersymmetry breaking, 191, 192, 209
dynamical supersymmetry breaking, (3,2) model, 210
dyons, 126
Dyson, F., 4

e^+e^- annihilation, 51, 53
effective action, string theory, 334
electric dipole moment, 68, 101
electric–magnetic duality, 128, 224
electroweak baryon number violation, 497
enhanced symmetries, 381
equilibrium conditions, 266
equivalence principle, 244
extended technicolor, 134

Faddeev–Popov ghosts, 5, 20
Fayet–Iliopoulos D term, 158
Fermi theory of weak interactions, 63
Fermi, E., 3, 63
fermions, compactified, 378
Feynman rules, Yang–Mills theories, 20
Feynman, R., 4, 57
finite temperature effective potential, 497
finite temperature field theory, 492, 493, 496
fixed points, 388
flat directions, 195
Friedmann equation, 261
FRW metric, 259, 260
FRW universe, $T(t)$, 265
FRW universe, scalar field equation, 265
Fujikawa, K., 80, 83
fundamental representation, 11

Gaillard, M., 6
gauge fixing, theories with spontaneous symmetry breaking, 20
gauge mediation, 214
gaugino condensation, 193
Gauss law constraint, 98
Gell-Mann, M., 102
general coordinate transformations, 245
geodesic equations, 248
Georgi, H., 107, 122
Glashow, S., 6, 107, 122
global symmetries, 168
gluino condensation, 208
goldstino, 148, 160
Goldstone's theorem, 487, 488
grand unified theories, 107
grand unified theories, mass relations, 112
grand unified theories, symmetry breaking, 111
grand unified theories, Weinberg angle, 109
Grassmann coordinates, 141
gravitational action, 250
gravitino problem, 279
Green's function, 2-d, 324
Green, M., 308
Green–Schwarz fermions, 353, 354
Greenberg, W., 6
Gross, D., 6
GSO projection, 348
GSO projection, modular invariance, 349
Guth, A., 278

Hamiltonian lattice, 50
Hawking radiation, 255
Hawking temperature, 255
Hawking, black hole puzzle, 306
Hawking, S., 256, 306
heavy quark potential, 39
heterotic string, $E_8 \times E_8$, 361
heterotic string, $O(16) \times O(16)$, 363
heterotic string, bosonic formulation, 387
heterotic string, interactions, 362
heterotic string, modular invariance, 363
heterotic string, O(32), 360
heterotic theory on Calabi–Yau manifolds, 418
hidden sector, 164
hierarchy problem, 71, 131
Higgs field, 24
Higgs mechanism, 16
Higgs particle, 5, 32
Higgs, P., 5
Hodge numbers, 405, 410
holography, 461
holomorphy, 152, 153, 208
holonomy, $SU(3)$, 407
Horava, P., 458

Horava–Witten theory, 458
hybrid inflation, 278
hypermultiplet, 220

IIA supergravity, 367
IIB self-duality, 456
IIB supergravity, 367
ILC, 475
Iliopoulos, J., 6
inflation, 272
inflation, flatness puzzle, 270
inflation, fluctuations, 275, 276
inflation, homogeneity puzzle, 270
inflation, reheating, 274
instanton, singular gauge, 203
instantons, 75, 86, 88, 89, 201
instantons, fermion zero modes, 93
instantons, QCD, 86
instantons, scale size integral, 94
instantons, spontaneously broken gauge theories, 96
instantons, symmetry violation, 95
Intriligator–Thomas models, 237

jets, 53

K–$\bar{K}$ mixing, 182
Kahler manifolds, 408
Kaluza, F., 373, 374
Kaluza–Klein theory, 373
kink, 121
Klein, O., 373, 374

landscape, 310
large N expansion, QCD, 99
large N limit, 85
lattice gauge theory, 44, 46
Lee, B., 6
leptogenesis, 290, 291, 292
lepton masses, 34
lepton number conservation, 63
LHC, 475
light cone gauge, 314, 315, 316, 317, 318
light element abundances, 266
lightest supersymmetric particle (LSP), 168
LSP, dark matter, 283

magnetic dipole moment, 68
magnetic monopole, 119
magnetic monopoles, 113
Maiani, L., 6
Maldacena, J., 463
matter dominated era, 262
metric tensor, 244
Mills, R., 5
Möbius group, 332
modular invariance, one-loop amplitudes, 337, 338
modular transformations, 336
moduli, 151, 195, 199, 375
moduli problem, 285
moduli spaces, 199
moduli spaces, approximate, 209
moduli, and inflation, 278
moduli, lifetime, 286
monopole problem, 270
MSSM, 167, 169
MSSM, breaking of $SU(2) \times U(1)$, 173
MSSM, Higgs mass limits, 176
MSSM, parameters, 171
mu term, 178

$N = 2$ theories, 219
$N = 2$, vector multiplet, 220
$N = 4$ Yang–Mills theory, 221
Nambu, Y., 307
Nambu–Goldstone bosons, 14
Ne'eman, Y., 102
Neumann boundary conditions, 316, 444
neutrino mass, 66
neutrino oscillations, 67
neutron electric dipole moment, 69
non-linear sigma model, 133
non-perturbative superpotential, 198
non-renormalization theorems, 151, 152, 153
non-renormalization theorems, string theory, 430, 431, 432, 433
normal ordering, 346, 347
NS 5-brane, 456
NS sector, 342
nucleosynthesis, 266

O'Raifeartaigh models, 157
$O(10)$, 114
one-loop potential, 161
OPE, 324
open strings, 315
open superstrings, 341
orbifolds, 388, 389, 390, 391
orbifolds, discrete symmetries, 393
orbifolds, discrete symmetries and flat directions, 394
orbifolds, effective actions, 396
orbifolds, modular invariance, 394
orbifolds, $N = 1$ supersymmetry, 390
orientifolds, 447

parity, 24
partons, 4
Pauli–Villars, 501
Pauli–Villars fields, 213
Pauli–Villars regulator, 78
Peccei, R., 103
Peccei–Quinn symmetry, 103
Penzias, A., 266
perimeter law, 49
pion masses, 488
pions as Goldstone bosons, 91
Politzer, D., 6
Prasad, M. K., 124
Prasad–Sommerfield monopole, 124
proton decay, 70, 114
proton decay, supersymmetric GUTs, 188

QCD at high temperature, infrared divergences, 496

QCD instantons, 92
QCD theta, 100, 101
QCD theta and massless u quark, 102
QCD theta and spontaneous CP violation, 103
QCD theta parameter, 71, 81, 89, 92
quadratic divergences, cancellations in supersymmetry, 171
quantum moduli space, 211
quark condensate, 90, 132
quark distribution functions, 58
quark mass matrix, 27
quark masses, 34
Quinn, H., 103

R-symmetries, 195
R-symmetry, 149, 170
R-parity, 168, 189
radiation dominated era, 262, 263
Ramond sector, 342, 343
Randall, R., 278
Randall–Sundrum, 475
Ricci scalar, 250
Ricci tensor, 249
right-handed neutrino, 116

S-matrix, factorization, 333
S-matrix, string theory, 331
Sakharov, A., 287
Salam, A., 5
Scherk, J., 308
Schwarzschild metric, 252, 253, 254
Schwarzschild radius, 254
Schwarz, J., 308
Schwinger, J., 4
Seiberg duality, 240
Seiberg, N., 152, 198, 207, 227, 230, 238, 240
Seiberg–Witten theory, 225, 226, 227, 228, 229, 230
Shenker, S., 462
Shifman, M., 501
SLAC, 4
slow roll approximation, 272
small gauge transformations, 98, 126
soft breaking, constraints, 181, 182
soft breaking, experimental constraints, 180
soft breakings, experimental constraints, 179
soft masses, renormalization group equations, 175
soft supersymmetry breaking, 162, 169
solitons, 120
Sommerfield, C., 124
sphaleron, 499
spin connection, 257
spinor representations of orthogonal groups, 115
spinors, general relativity, 256
spontaneous supersymmetry breaking, 157
stress tensor, 250
string coupling, 368
string theory, Fayet–Iliopoulos terms, 434, 435, 436, 437
string theory, finiteness, 306
string theory, gaugino condensation, 437, 438
string theory, unification of couplings, 416, 424
strings, background fields, 382, 383
strong CP problem, 76
strongly coupled heterotic string theory, 458, 459
$SU(5)$, 107, 108
superconformal zero modes, 204
superfields, 142
supergravity, eleven dimensions, 365, 366
superpotential, 145
superspace, 141
superspace covariant derivatives, 142
superstring action, 341
superstring mode expansion, 342
superstrings, space-time fermions, 343
superstrings, vertex operators, 355
supersymmetric guts, 185
supersymmetric qcd, 195
supersymmetry, 475
supersymmetry algebra, 140
supersymmetry breaking, early universe, 296
supersymmetry breaking, supergravity, 164
supersymmetry breaking, vanishing of the ground state energy, 148
supersymmetry currents, 147
supersymmetry generators, 142
supersymmetry representations, 140
supersymmetry zero modes, 204
supersymmetry, component Lagrangian, 146
supersymmetry, world sheet, 345
Susskind, L., 307, 462
Susy QCD, $N_{\rm f} < N - 1$, 200
Susy QCD, $N_{\rm f} = N - 1$, 201

't Hooft, G., 5, 40, 94, 96, 201, 238
't Hooft–Polyakov monopole, 124
T-duality, 381
T-duality, open strings and D-branes, 449
technicolor, 131, 475
theta functions, 350
Tomanaga, S., 4
top quark, 39
top quark, symmetry breaking in MSSM, 175
toroidal compactification, 386
toroidal compactification, momentum lattice, 379
toroidal compactification, non-supersymmetric, 398
Type I-$O(32)$ duality, 457
Type II strings, spectra, 347

$U(1)$ problem, 99
unification of couplings, 110
unitarity triangle, 33

Vainshtein, A., 501
vector superfields, 143
Veltman, M., 5, 40
Veneziano amplitude, 333
Veneziano, G., 307
vertex operators, 329, 330
vielbein, 256, 257
Virasoro algebra, 327
Virasoro–Shapiro amplitude, 333

visible sector, 164
vortices, 122

W boson, 135
W bosons, 26, 29, 133
Weinberg, S., 5, 103, 479
Wess–Zumino model, 149
Weyl basis, 483, 484
Weyl rescaling, 375
Wilczek, W., 6
Wilson line, 46
Wilson lines on Calabi–Yau manifolds, 420
Wilson lines, compactified, 379
Wilson loop, 47
Wilson, K., 36, 46, 71
Wilson, Ken, 47
Wilson, R., 266
winding modes, 377
Witten effect, 127
Witten index, 192
Witten, E., 127, 192, 227, 230, 308, 458
Wolfenstein parameterization, 33

Yang, C. N., 5
Yang–Mills action, 10
Yang–Mills theory, 9
Yoneya, T., 308
Yukawa, H., 3

Z boson, 26, 135
Z bosons, 29, 133
Zaks, A., 234

A First Course in String Theory

Contents

Foreword by David Gross *page* xiii
From the Preface to the First Edition xv
Preface to the Second Edition xix

Part I Basics 1

1 A brief introduction 3
1.1 The road to unification 3
1.2 String theory as a unified theory of physics 6
1.3 String theory and its verification 9
1.4 Developments and outlook 11

2 Special relativity and extra dimensions 13
2.1 Units and parameters 13
2.2 Intervals and Lorentz transformations 15
2.3 Light-cone coordinates 22
2.4 Relativistic energy and momentum 26
2.5 Light-cone energy and momentum 28
2.6 Lorentz invariance with extra dimensions 30
2.7 Compact extra dimensions 31
2.8 Orbifolds 35
2.9 Quantum mechanics and the square well 36
2.10 Square well with an extra dimension 38

3 Electromagnetism and gravitation in various dimensions 45
3.1 Classical electrodynamics 45
3.2 Electromagnetism in three dimensions 47
3.3 Manifestly relativistic electrodynamics 48
3.4 An aside on spheres in higher dimensions 52
3.5 Electric fields in higher dimensions 55
3.6 Gravitation and Planck's length 58
3.7 Gravitational potentials 62
3.8 The Planck length in various dimensions 63
3.9 Gravitational constants and compactification 64
3.10 Large extra dimensions 67

4 Nonrelativistic strings 73
- 4.1 Equations of motion for transverse oscillations 73
- 4.2 Boundary conditions and initial conditions 75
- 4.3 Frequencies of transverse oscillation 76
- 4.4 More general oscillating strings 77
- 4.5 A brief review of Lagrangian mechanics 78
- 4.6 The nonrelativistic string Lagrangian 81

5 The relativistic point particle 89
- 5.1 Action for a relativistic point particle 89
- 5.2 Reparameterization invariance 93
- 5.3 Equations of motion 94
- 5.4 Relativistic particle with electric charge 97

6 Relativistic strings 100
- 6.1 Area functional for spatial surfaces 100
- 6.2 Reparameterization invariance of the area 103
- 6.3 Area functional for spacetime surfaces 106
- 6.4 The Nambu-Goto string action 111
- 6.5 Equations of motion, boundary conditions, and D-branes 112
- 6.6 The static gauge 116
- 6.7 Tension and energy of a stretched string 118
- 6.8 Action in terms of transverse velocity 120
- 6.9 Motion of open string endpoints 124

7 String parameterization and classical motion 130
- 7.1 Choosing a σ parameterization 130
- 7.2 Physical interpretation of the string equation of motion 132
- 7.3 Wave equation and constraints 134
- 7.4 General motion of an open string 136
- 7.5 Motion of closed strings and cusps 142
- 7.6 Cosmic strings 145

8 World-sheet currents 154
- 8.1 Electric charge conservation 154
- 8.2 Conserved charges from Lagrangian symmetries 155
- 8.3 Conserved currents on the world-sheet 159
- 8.4 The complete momentum current 161
- 8.5 Lorentz symmetry and associated currents 165
- 8.6 The slope parameter α' 168

9 Light-cone relativistic strings 175
- 9.1 A class of choices for τ 175
- 9.2 The associated σ parameterization 178

9.3 Constraints and wave equations 182
9.4 Wave equation and mode expansions 183
9.5 Light-cone solution of equations of motion 186

10 Light-cone fields and particles 194
10.1 Introduction 194
10.2 An action for scalar fields 195
10.3 Classical plane-wave solutions 197
10.4 Quantum scalar fields and particle states 200
10.5 Maxwell fields and photon states 206
10.6 Gravitational fields and graviton states 209

11 The relativistic quantum point particle 216
11.1 Light-cone point particle 216
11.2 Heisenberg and Schrödinger pictures 218
11.3 Quantization of the point particle 220
11.4 Quantum particle and scalar particles 225
11.5 Light-cone momentum operators 226
11.6 Light-cone Lorentz generators 229

12 Relativistic quantum open strings 236
12.1 Light-cone Hamiltonian and commutators 236
12.2 Commutation relations for oscillators 241
12.3 Strings as harmonic oscillators 246
12.4 Transverse Virasoro operators 250
12.5 Lorentz generators 259
12.6 Constructing the state space 262
12.7 Equations of motion 268
12.8 Tachyons and D-brane decay 270

13 Relativistic quantum closed strings 280
13.1 Mode expansions and commutation relations 280
13.2 Closed string Virasoro operators 286
13.3 Closed string state space 290
13.4 String coupling and the dilaton 294
13.5 Closed strings on the $\mathbb{R}^1/\mathbb{Z}_2$ orbifold 296
13.6 The twisted sector of the orbifold 298

14 A look at relativistic superstrings 307
14.1 Introduction 307
14.2 Anticommuting variables and operators 308
14.3 World-sheet fermions 309
14.4 Neveu−Schwarz sector 312
14.5 Ramond sector 315

14.6 Counting states 317
14.7 Open superstrings 320
14.8 Closed string theories 322

Part II Developments 329

15 D-branes and gauge fields 331
15.1 Dp-branes and boundary conditions 331
15.2 Quantizing open strings on Dp-branes 333
15.3 Open strings between parallel Dp-branes 338
15.4 Strings between parallel Dp- and Dq-branes 345

16 String charge and electric charge 356
16.1 Fundamental string charge 356
16.2 Visualizing string charge 362
16.3 Strings ending on D-branes 365
16.4 D-brane charges 370

17 T-duality of closed strings 376
17.1 Duality symmetries and Hamiltonians 376
17.2 Winding closed strings 378
17.3 Left movers and right movers 381
17.4 Quantization and commutation relations 383
17.5 Constraint and mass formula 386
17.6 State space of compactified closed strings 388
17.7 A striking spectrum coincidence 392
17.8 Duality as a full quantum symmetry 394

18 T-duality of open strings 400
18.1 T-duality and D-branes 400
18.2 $U(1)$ gauge transformations 404
18.3 Wilson lines on circles 406
18.4 Open strings and Wilson lines 410

19 Electromagnetic fields on D-branes 415
19.1 Maxwell fields coupling to open strings 415
19.2 D-branes with electric fields 418
19.3 D-branes with magnetic fields 423

20 Nonlinear and Born–Infeld electrodynamics 433
20.1 The framework of nonlinear electrodynamics 433
20.2 Born–Infeld electrodynamics 438
20.3 Born–Infeld theory and T-duality 443

21 String theory and particle physics 451
21.1 Intersecting D6-branes 451
21.2 D-branes and the Standard Model gauge group 457
21.3 Open strings and the Standard Model fermions 463
21.4 The Standard Model on intersecting D6-branes 472
21.5 String theory models of particle physics 479
21.6 Moduli stabilization and the landscape 481

22 String thermodynamics and black holes 495
22.1 A review of statistical mechanics 495
22.2 Partitions and the quantum violin string 498
22.3 Hagedorn temperature 505
22.4 Relativistic particle partition function 507
22.5 Single string partition function 509
22.6 Black holes and entropy 513
22.7 Counting states of a black hole 517

23 Strong interactions and AdS/CFT 525
23.1 Introduction 525
23.2 Mesons and quantum rotating strings 526
23.3 The energy of a stretched effective string 531
23.4 A large-N limit of a gauge theory 533
23.5 Gravitational effects of massive sources 535
23.6 Motivating the AdS/CFT correspondence 537
23.7 Parameters in the AdS/CFT correspondence 541
23.8 Hyperbolic spaces and conformal boundary 543
23.9 Geometry of AdS and holography 549
23.10 AdS/CFT at finite temperature 554
23.11 The quark–gluon plasma 559

24 Covariant string quantization 568
24.1 Introduction 568
24.2 Open string Virasoro operators 570
24.3 Selecting the quantum constraints 572
24.4 Lorentz covariant state space 577
24.5 Closed string Virasoro operators 580
24.6 The Polyakov string action 582

25 String interactions and Riemann surfaces 591
25.1 Introduction 591
25.2 Interactions and observables 592
25.3 String interactions and global world-sheets 595
25.4 World-sheets as Riemann surfaces 598
25.5 Schwarz–Christoffel map and three-string interaction 602

25.6 Moduli spaces of Riemann surfaces 608
25.7 Four open string interaction 617
25.8 Veneziano amplitude 622

26 Loop amplitudes in string theory 630
26.1 Loop diagrams and ultraviolet divergences 630
26.2 Annuli and one-loop open strings 631
26.3 Annuli and electrostatic capacitance 636
26.4 Non-planar open string diagrams 642
26.5 Four closed string interactions 643
26.6 The moduli space of tori 646

References 659
Index 667

Foreword

String theory is one of the most exciting fields in theoretical physics. This ambitious and speculative theory offers the potential of unifying gravity and all the other forces of nature and all forms of matter into one unified conceptual structure.

String theory has the unfortunate reputation of being impossibly difficult to understand. To some extent this is because, even to its practitioners, the theory is so new and so ill understood. However, the basic concepts of string theory are quite simple and should be accessible to students of physics with only advanced undergraduate training.

I have often been asked by students and by fellow physicists to recommend an introduction to the basics of string theory. Until now all I could do was point them either to popular science accounts or to advanced textbooks. But now I can recommend to them Barton Zwiebach's excellent book.

Zwiebach is an accomplished string theorist, who has made many important contributions to the theory, especially to the development of string field theory. In this book he presents a remarkably comprehensive description of string theory that starts at the beginning, assumes only minimal knowledge of advanced physics, and proceeds to the current frontiers of physics. Already tested in the form of a very successful undergraduate course at MIT, Zwiebach's exposition proves that string theory can be understood and appreciated by a wide audience.

I strongly recommend this book to anyone who wants to learn the basics of string theory.

David Gross
Director, Kavli Institute For Theoretical Physics
University of California, Santa Barbara

From the Preface to the First Edition

The idea of having a serious string theory course for undergraduates was first suggested to me by a group of MIT sophomores sometime in May of 2001. I was teaching Statistical Physics, and I had spent an hour-long recitation explaining how a relativistic string at high energies appears to approach a constant temperature (the Hagedorn temperature). I was intrigued by the idea of a basic string theory course, but it was not immediately clear to me that a useful one could be devised at this level.

A few months later, I had a conversation with Marc Kastner, the Physics Department Head. In passing, I told him about the sophomores' request for a string theory course. Kastner's instantaneous and enthusiastic reaction made me consider seriously the idea for the first time. At the end of 2001, a new course was added to the undergraduate physics curriculum at MIT. In the spring term of 2002 I taught *String Theory for Undergraduates* for the first time. This book grew out of the lecture notes for that course.

When we think about teaching string theory at the undergraduate level the main question is, "Can the material really be explained at this level?". After teaching the subject two times, I am convinced that the answer to the question is a definite yes. Although a complete mastery of string theory requires a graduate-level physics education, the basics of string theory can be well understood with the limited tools acquired in the first two or three years of an undergraduate education.

What is the value of learning string theory, for an undergraduate? By exposing the students to cutting-edge ideas, a course in string theory can help nurture the excitement and enthusiasm that led them to choose physics as a major. Moreover, students will find in string theory an opportunity to sharpen and refine their understanding of most of the undergraduate physics curriculum. This is valuable even for students who do not plan to specialize in theoretical physics.

This book was tailored to be understandable to an advanced undergraduate. Therefore, I believe it will be a readable introduction to string theory for any graduate student or, in fact, for any physicist who wants to learn the basics of string theory.

Acknowledgements

I would like to thank Marc Kastner, Physics Department Head, for his enthusiastic support and his interest. I am also grateful to Thomas Greytak, Associate Head for Education, and to Robert Jaffe, Director of the Center for Theoretical Physics, both of whom kindly supported this project.

Teaching string theory to a class composed largely of bright undergraduates was both a stimulating and a rewarding experience. I am grateful to the group of students that composed the first class:

Jeffrey Brock	Adam Granich	Trisha Montalbo
Zilong Chen	Markéta Havlíčková	Eugene Motoyama
Blair Connely	Kenneth Jensen	Megha Padi
Ivailo Dimov	Michael Krypel	Ian Parrish
Peter Eckley	Francis Lam	James Pate
Qudsia Ejaz	Philippe Larochelle	Timothy Richards
Kasey Ensslin	Gabrielle Magro	James Smith
Teresa Fazio	Sourav Mandal	Morgan Sonderegger
Caglar Girit	Stefanos Marnerides	David Starr
Donglai Gong		

They were enthusiastic, funny, and lively. My lectures were voice-recorded and three of the students, Gabrielle Magro, Megha Padi, and David Starr, turned the tapes and the blackboard equations into LaTeX files. I am grateful to the three of them for their dedication and for the care they took in creating accurate files. They provided the impetus to start the process of writing a book. I edited the files to produce lecture notes.

Additional files for a set of summer lectures were created by Gabrielle and Megha. In the next six months the lecture notes became the draft for a book. After teaching the course for a second time in the spring term of 2003 and a long summer of edits and revisions, the book was completed in October 2003.

By the time the lecture notes had become a book draft, David Starr offered to read it critically. He basically marked every paragraph, suggesting improvements in the exposition and demonstrating an uncanny ability to spot weak points. His criticism forced me to go through major rewriting. His input was tremendous. Whatever degree of clarity has been achieved, it is in no small measure thanks to his effort.

I am delighted to acknowledge help and advice from my friend and colleague Jeffrey Goldstone. He shared generously his understanding of string theory, and several sections in this book literally grew out of his comments. He helped me teach the course the second time that it was offered. While doing so, he offered perceptive criticism of the whole text. He also helped improve many of the problems, for which he wrote elegant solutions.

The input of my friend and collaborator Ashoke Sen was critical. He believed that string theory could be taught at a basic level and encouraged me to try to do it. I consulted repeatedly with him about the topics to be covered and about the strategies to present them. He kindly read the first full set of lecture notes and gave invaluable advice that helped shape the form of this book.

The help and interest of many people made writing this book a very pleasant task. For detailed comments on all of its content I am indebted to Chien-Hao Liu and to James Stasheff. Alan Dunn and Blake Stacey helped test the problems that could not be assigned in class. Jan Troost was a sounding board and provided advice and criticism. I've relied on the knowledge of my string theory colleagues – Amihay Hanany, Daniel Freedman, and Washington Taylor. I'd like to thank Philip Argyres, Andreas Karch, and Frieder Lenz

for testing the lecture notes with their students. Juan Maldacena and Samir Mathur provided helpful input on the subject of string thermodynamics and black holes. Boris Körs, Fernando Quevedo, and Angel Uranga helped and advised on the subject of string phenomenology. Thanks are also due to Tamsin van Essen, editor at Cambridge, for her advice and her careful work during the entire publishing process.

Finally, I would like to thank my wife Gaby and my children Cecile, Evy, Margaret, and Aaron. At every step of the way I was showered with their love and support. Cecile and Evy read parts of the manuscript and advised on language. Questions on string theory from Gaby and Margaret tested my ability to explain. Young Aaron insisted that a ghost sitting on a string would make a perfect cover page, but we settled for strings moving in electric fields.

Barton Zwiebach

Cambridge, Massachusetts, 2003

Preface to the Second Edition

It has been almost five years since I finished writing the first edition of *A First Course in String Theory*. I have since taught the undergraduate string theory course at MIT three times, and I have received comments and suggestions from colleagues all over the world. I have learned what parts of the book are most challenging for the students, and I have heard requests for extra material.

As in the first edition, the book is broadly divided into Part I (Basics) and Part II (Developments). In this second edition I have improved the clarity of many arguments and the general readability of Part I. This part is studied by the largest number of readers, many of them independently and outside of the classroom setting. The changes should make study easier. There are more figures and the number of problems has been increased to better cover the range of ideas developed in the text. Part I has five new sections and one new chapter. The new sections discuss the classical motion of closed strings, cosmic strings, and orbifolds. The new chapter, Chapter 14, is the last one of Part I. It explains the basics of superstring theory.

Part II has changed as well. The ordering of chapters has been altered to bring T-duality earlier into the book. The material relevant to particle physics has been collected in Chapter 21 and includes a new section on moduli stabilization and the landscape. Chapter 23 is new and is entirely devoted to strong interactions and the AdS/CFT correspondence. I aim to give there a gentle introduction to this lively area of research. The number of chapters in the book has gone from twenty-three to twenty-six, a nice number to end a book on string theory!

I want to thank Hong Liu and Juan Maldacena for helpful input on the subject of AdS/CFT. Many thanks are also due to Alan Guth, who helped me teach the string theory course in the spring term of 2007. He tested many of the new problems and offered very valuable criticism of the text.

About this book

A First Course in String Theory should be accessible to anyone who has been exposed to special relativity, basic quantum mechanics, electromagnetism, and introductory statistical physics. Some familiarity with Lagrangian mechanics is useful but not indispensable.

Except for the introduction, all chapters contain exercises and problems. The exercises, called *Quick calculations*, are inserted at various points throughout the text. They are control calculations that are expected to be straightforward. Undue difficulty in carrying them out may indicate problems understanding the material. The problems at the end of the

chapters are more challenging and sometimes develop new ideas. A problem marked with a dagger† is one whose results are cited later in the text. A mastery of the material requires solving all the exercises and many of the problems. All the problems should be read, at least.

Throughout most of the book the material is developed in a self-contained way, and very little must be taken on faith. Chapters 14, 21, 22, and 23 contain a few sections that address subjects of much interest for which a full explanation cannot be provided at the level of this book. The reader will be asked to accept some reasonable facts at face value, but otherwise the material is developed logically and should be *fully* understandable. These sections are *not* addressed to experts.

This book has two parts. Part I is called "Basics," and Part II is called "Developments." Part I begins with Chapter 1 and concludes with Chapter 14. Part II comprises the rest of the book: it begins with Chapter 15 and it ends with Chapter 26.

Chapter 1 serves as an introduction. Chapter 2 reviews special relativity, but it also introduces concepts that are likely to be new: light-cone coordinates, light-cone energy, compact extra dimensions, and orbifolds. In Chapter 3 we review electrodynamics and its manifestly relativistic formulation. We make some comments on general relativity and study the effect of compact dimensions on the Planck length. We are able at this point to examine the exciting possibility that large extra dimensions may exist. Chapter 4 uses nonrelativistic strings to develop some intuition, to review the Lagrangian formulation of mechanics, and to introduce terminology. Chapter 5 uses the relativistic point particle to prepare the ground for the study of the relativistic string. The power and elegance of the Lagrangian formulation become evident at this point. The first encounter with string theory happens in Chapter 6, which deals with the classical dynamics of the relativistic string. This is a very important chapter, and it must be understood thoroughly. Chapter 7 solidifies the understanding of string dynamics through the detailed study of string motion, both for open and for closed strings. It includes a section on cosmic strings, a topic of potential experimental relevance. Chapters 1 through 7 could comprise a mini-course in string theory.

Chapters 8 through 11 prepare the ground for the quantization of relativistic strings. In Chapter 8, one learns how to calculate conserved quantities, such as the momentum and the angular momentum of free strings. Chapter 9 gives the light-cone gauge solution of the string equations of motion and introduces the terminology that is used in the quantum theory. Chapter 10 explains the basics of quantum fields and particle states, with emphasis on the counting of the parameters that characterize scalar field states, photon states, and graviton states. In Chapter 11 we perform the light-cone gauge quantization of the relativistic particle. It all comes together in Chapter 12, another important chapter that should be understood thoroughly. This chapter presents the light-cone gauge quantization of the open relativistic string. The critical dimension is obtained and photon states are shown to emerge. Chapter 12 contains a section on the subject of tachyon condensation. Chapter 13 discusses the quantization of closed strings and the emergence of graviton states. It also contains two sections that deal with quantum closed strings on the simplest orbifold, the half-line. Chapter 14 is the last chapter of Part I. It introduces the subject of superstrings. The Ramond and Neveu–Schwarz sectors of open strings are presented and combined to

obtain a supersymmetric theory. The chapter concludes with a brief discussion of type II closed string theories.

The first part of this book can be characterized as an uphill road that leads to the quantization of the string at the summit. In the second part of this book the climb is over. The pace slows down a little, and the material elaborates upon previously introduced ideas. In Part II one reaps many rewards for the effort exerted in Part I.

The first chapter in Part II, Chapter 15, deals with the important subject of open strings on various D-brane configurations. The discussion of orientifolds has been relegated to the problems at the end of the chapter. Chapter 16 introduces the concept of string charge and demonstrates that the endpoints of open strings carry Maxwell charge. The next four chapters are organized around the fascinating subject of T-duality. Chapters 17 and 18 present the T-duality properties of closed and open strings, respectively. Chapter 19 studies D-branes with electromagnetic fields, using T-duality as the main tool. Chapter 20 introduces the general framework of nonlinear electrodynamics. It demonstrates that electromagnetic fields in string theory are governed by Born–Infeld theory, a nonlinear theory in which the self-energy of point charges is finite.

String models of particle physics are considered in Chapter 21. This chapter explains in detail the particle content of the Standard Model and discusses one approach, based on intersecting D6-branes, to the construction of a realistic string model. The chapter concludes with some material on moduli stabilization and the landscape.

Chapter 22 begins with string thermodynamics, followed by the subject of black hole entropy. It presents string theory attempts to derive the entropy of Schwarzschild black holes and the successful derivation of the entropy for a supersymmetric black hole. The applications of string theory to strong interactions are studied in Chapter 23. After a discussion of Regge trajectories and the quark–antiquark potential, the subject turns to the AdS/CFT correspondence. The correspondence is discussed in some detail, with emphasis on the geometry of AdS spaces. A section on the quark–gluon plasma is included.

Chapter 24 gives an introduction to the Lorentz covariant quantization of strings. It also introduces the Polyakov string action. The last two chapters in the book, Chapters 25 and 26, examine string interactions. We learn that the string diagrams which represent the processes of string interactions are Riemann surfaces. These two chapters assume a little familiarity with complex variables and have a mathematical flavor. One important goal here is to provide insight into the absence of ultraviolet divergences in string theory, the fact that made string theory the first candidate for a theory of quantum gravity.

In this book I have tried to emphasize the connections with ideas that students have learned before. The quantization of strings is described as the quantization of an infinite number of oscillators. String charge is visualized as a Maxwell current. The effects of Wilson lines on circles are compared with the Bohm–Aharonov effect. The modulus of an annulus is related to the capacitance of a cylindrical conductor, and so forth and so on. The treatment of topics is generally explicit and detailed, with formalism kept to a minimum.

The choice was made to use the light-cone gauge to quantize the strings. This approach to quantization can be understood in full detail by students with some prior exposure to

quantum mechanics. The same is *not* true for the Lorentz covariant quantization of strings, where states of negative norms must be dealt with, the Hamiltonian vanishes, and there is no conventional looking Schrödinger equation. The light-cone approach suffices for most physical problems and, in fact, simplifies the treatment of several questions.

This book as a textbook

Part I of the book is structured tightly. Little can be omitted without hampering the understanding of string quantization. The first chapter in Part II (on D-branes) is important for much of the later material. Many choices among the remaining chapters are possible. Different readers/instructors may take different routes.

My experience suggests that the complete book can be covered in a full-year course at the undergraduate level. In a school with an academic year composed of three quarters, Part I and four chapters from Part II may be covered in two quarters. In a school with an academic year composed of two semesters, Part I and two chapters from Part II may be covered in one semester. In either case, the choice of chapters from Part II is a matter of taste. Chapters 21, 22, and 23 give an appreciation for current research in string theory. Lecturers who prefer to focus on T-duality and its implications will cover as much as possible from Chapters 17–20. If this book is used to teach exclusively to graduate students, the pace can be quickened considerably.

An updated list of corrections can be found at http://xserver.lns.mit.edu/∼zwiebach/firstcourse.html. Solutions to the problems in the book are available to lecturers via solutions@cambridge.org.

PART I

BASICS

1 A brief introduction

Here we meet string theory for the first time. We see how it fits into the historical development of physics, and how it aims to provide a unified description of all fundamental interactions.

1.1 The road to unification

Over the course of time, the development of physics has been marked by unifications: events when different phenomena were recognized to be related and theories were adjusted to reflect such recognition. One of the most significant of these unifications occurred in the nineteenth century.

For a while, electricity and magnetism had appeared to be unrelated physical phenomena. Electricity was studied first. The remarkable experiments of Henry Cavendish were performed in the period from 1771 to 1773. They were followed by the investigations of Charles Augustin de Coulomb, which were completed in 1785. These works provided a theory of static electricity, or electrostatics. Subsequent research into magnetism, however, began to reveal connections with electricity. In 1819 Hans Christian Oersted discovered that the electric current on a wire can deflect the needle of a compass placed nearby. Shortly thereafter, Jean-Baptiste Biot and Felix Savart (1820) and André-Marie Ampère (1820–1825) established the rules by which electric currents produce magnetic fields. A crucial step was taken by Michael Faraday (1831), who showed that changing magnetic fields generate electric fields. Equations that described all of these results became available, but they were, in fact, inconsistent. It was James Clerk Maxwell (1865) who constructed a consistent set of equations by adding a new term to one of the equations. Not only did this term remove the inconsistencies, but it also resulted in the prediction of electromagnetic waves. For this great insight, the equations of *electromagnetism* (or electrodynamics) are now called "Maxwell's equations." These equations unify electricity and magnetism into a consistent whole. This elegant and aesthetically pleasing unification was not optional. Separate theories of electricity and magnetism would be inconsistent.

Another fundamental unification of two types of phenomena occurred in the late 1960s, about one-hundred years after the work of Maxwell. This unification revealed the deep relationship between electromagnetic forces and the forces responsible for weak interactions. To appreciate the significance of this unification it is necessary first to review the main developments that occurred in physics since the time of Maxwell.

An important change of paradigm was triggered by Albert Einstein's special theory of relativity. In this theory one finds a striking conceptual unification of the separate notions of space and time. Different from a unification of forces, the merging of space and time into a spacetime continuum represented a new recognition of the nature of the *arena* where physical phenomena take place. Newtonian mechanics was replaced by relativistic mechanics, and older ideas of absolute time were abandoned. Mass and energy were shown to be interchangeable.

Another change of paradigm, perhaps an even more dramatic one, was brought forth by the discovery of quantum mechanics. Developed by Erwin Schrödinger, Werner Heisenberg, Paul Dirac and others, quantum theory was verified to be the correct framework to describe microscopic phenomena. In quantum mechanics classical observables become operators. If two operators fail to commute, the corresponding observables cannot be measured simultaneously. Quantum mechanics is a framework, more than a theory. It gives the rules by which theories must be used to extract physical predictions.

In addition to these developments, four fundamental forces had been recognized to exist in nature. Let us have a brief look at them.

One of them is the force of gravity. This force has been known since antiquity, but it was first described accurately by Isaac Newton. Gravity underwent a profound reformulation in Albert Einstein's theory of general relativity. In this theory, the spacetime arena of special relativity acquires a life of its own, and gravitational forces arise from the curvature of this dynamical spacetime. Einstein's general relativity is a classical theory of gravitation. It is not formulated as a quantum theory.

The second fundamental force is the electromagnetic force. As we discussed above, the electromagnetic force is well described by Maxwell's equations. Electromagnetism, or Maxwell theory, is formulated as a classical theory of electromagnetic fields. As opposed to Newtonian mechanics, which was modified by special relativity, Maxwell theory is fully consistent with special relativity.

The third fundamental force is the weak force. This force is responsible for the process of nuclear beta decay, in which a neutron decays into a proton, an electron, and an antineutrino. In general, processes that involve neutrinos are mediated by weak forces. While nuclear beta decay had been known since the end of the nineteenth century, the recognition that a new force was at play did not take hold until the middle of the twentieth century. The strength of this force is measured by the Fermi constant. Weak interactions are much weaker than electromagnetic interactions.

Finally, the fourth force is the strong force, nowadays called the color force. This force is at play in holding together the constituents of the neutron, the proton, the pions, and many other subnuclear particles. These constituents, called quarks, are held so tightly by the color force that they cannot be seen in isolation.

We are now in a position to return to the subject of unification. In the late 1960s the Weinberg–Salam model of *electroweak* interactions put together electromagnetism and the weak force into a unified framework. This unified model was neither dictated nor justified only by considerations of simplicity or elegance. It was necessary for a predictive and consistent theory of the weak interactions. The theory is initially formulated with four massless particles that carry the forces. A process of symmetry breaking gives mass to three of these

particles: the W^+, the W^-, and the Z^0. These particles are the carriers of the weak force. The particle that remains massless is the photon, which is the carrier of the electromagnetic force.

Maxwell's equations, as we discussed before, are equations of classical electromagnetism. They do not provide a quantum theory. Physicists have discovered quantization methods, which can be used to turn a classical theory into a quantum theory – a theory that can be calculated using the principles of quantum mechanics. While classical electrodynamics can be used confidently to calculate the transmission of energy in power lines and the radiation patterns of radio antennas, it is neither an accurate nor a correct theory for microscopic phenomena. Quantum electrodynamics (QED), the quantum version of classical electrodynamics, is required for correct computations in this arena. In QED, the photon appears as the quantum of the electromagnetic field. The theory of weak interactions is also a quantum theory of particles, so the correct, unified theory is the quantum electroweak theory.

The quantization procedure is also successful in the case of the strong color force, and the resulting theory has been called quantum chromodynamics (QCD). The carriers of the color force are eight massless particles. These are colored gluons, and just like the quarks, they cannot be observed in isolation. The quarks respond to the gluons because they carry color. Quarks can come in three colors.

The electroweak theory together with QCD form the Standard Model of particle physics. In the Standard Model there is some interplay between the electroweak sector and the QCD sector because some particles feel both types of forces. But there is no real and deep unification of the weak force and the color force. The Standard Model summarizes completely the present knowledge of particle physics. So, in fact, we are not certain about any possible further unification.

In the Standard Model there are twelve force carriers: the eight gluons, the W^+, the W^-, the Z^0, and the photon. All of these are bosons. There are also many matter particles, all of which are fermions. The matter particles are of two types: leptons and quarks. The leptons include the electron e^-, the muon μ^-, the tau τ^-, and the associated neutrinos ν_e, ν_μ, and ν_τ. We can list them as

$$\text{leptons:}\quad e^-, \mu^-, \tau^-, \nu_e, \nu_\mu, \nu_\tau.$$

Since we must include their antiparticles, this adds up to a total of twelve leptons. The quarks carry color charge, electric charge, and can respond to the weak force as well. There are six different types of quarks. Poetically called flavors, these types are: up (u), down (d), charm (c), strange (s), top (t), and bottom (b). We can list them as

$$\text{quarks:}\quad u, d, c, s, t, b.$$

The u and d quarks, for example, carry different electric charges and respond differently to the weak force. Each of the six quark flavors listed above comes in three colors, so this gives $6 \times 3 = 18$ particles. Including the antiparticles, we get a total of 36 quarks. Adding leptons and quarks together we have a grand total of 48 matter particles. Adding matter particles and force carriers together we have a total of 60 particles in the Standard Model.

Despite the large number of particles it describes, the Standard Model is reasonably elegant and very powerful. As a complete theory of physics, however, it has two significant

shortcomings. The first one is that it does not include gravity. The second one is that it has about twenty parameters that cannot be calculated within its framework. Perhaps the simplest example of such a parameter is the dimensionless (or unit-less) ratio of the mass of the muon to the mass of the electron. The value of this ratio is about 207, and it must be put into the model by hand.

Most physicists believe that the Standard Model is only a step towards the formulation of a complete theory of physics. A large number of physicists also suspect that some unification of the electroweak and strong forces into a Grand Unified Theory (GUT) will prove to be correct. At present, however, the unification of these two forces appears to be optional.

Another attractive possibility is that a more complete version of the Standard Model includes supersymmetry. Supersymmetry is a symmetry that relates bosons to fermions. Since all matter particles are fermions and all force carriers are bosons, this remarkable symmetry unifies matter and forces. In a theory with supersymmetry, bosons and fermions appear in pairs of equal mass. The particles of the Standard Model do not have this property, so supersymmetry, if it exists in nature, must be spontaneously broken. Supersymmetry is such an appealing symmetry that many physicists believe that it will eventually be discovered.

While the above extensions of the Standard Model may or may not occur, it is clear that the inclusion of gravity into the particle physics framework is not optional. Gravity must be included, with or without unification, if one is to have a complete theory. The effects of the gravitational force are presently quite negligible at the microscopic level, but they are crucial in studies of cosmology of the early universe.

There is, however, a major problem when one attempts to incorporate gravitational physics into the Standard Model. The Standard Model is a quantum theory, while Einstein's general relativity is a classical theory. It seems very difficult, if not altogether impossible, to have a consistent theory that is partly quantum and partly classical. Given the successes of quantum theory, it is widely believed that gravity must be turned into a quantum theory. The procedures of quantization, however, encounter profound difficulties in the case of gravity. The resulting theory of quantum gravity appears to be ill-defined. As a practical matter, in many circumstances one can work confidently with classical gravity coupled to the Standard Model. For example, this is done routinely in present-day descriptions of the universe. A theory of quantum gravity is necessary, however, to study physics at times very near to the Big Bang, and to study certain properties of black holes. Formulating a quantum theory that includes both gravity and the other forces seems fundamentally necessary. A *unification* of gravity with the other forces might be required to construct this complete theory.

1.2 String theory as a unified theory of physics

String theory is an excellent candidate for a unified theory of all forces in nature. It is also a rather impressive prototype of a complete theory of physics. In string theory all forces are truly unified in a deep and significant way. In fact, all the particles are unified. String

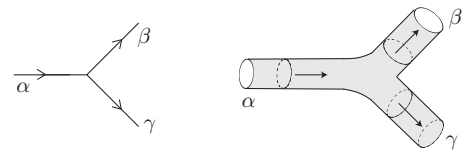

Fig. 1.1 The decay $\alpha \to \beta + \gamma$ as a particle process (left) and as a string process (right).

theory is a quantum theory, and, because it includes gravitation, it is a quantum theory of gravity. Viewed from this perspective, and recalling the failure of Einstein's gravity to yield a quantum theory, one may conclude that in string theory all other interactions are necessary for the consistency of the quantum gravitational sector! While it may be difficult to measure the effects of quantum gravity directly, a theory of quantum gravity such as string theory may have testable predictions concerning the other interactions.

Why is string theory a truly unified theory? The reason is simple and goes to the heart of the theory. In string theory, each particle is identified as a particular vibrational mode of an elementary microscopic string. A musical analogy is very apt. Just as a violin string can vibrate in different modes and each mode corresponds to a different sound, the modes of vibration of a fundamental string can be recognized as the different particles we know. One of the vibrational states of strings is the graviton, the quantum of the gravitational field. Since there is just one type of string, and all particles arise from string vibrations, all particles are naturally incorporated into a single theory. When we think in string theory of a decay process $\alpha \to \beta + \gamma$, where an elementary particle α decays into particles β and γ, we imagine a single string vibrating in such a way that it is identified as particle α that breaks into two strings that vibrate in ways that identify them as particles β and γ (Figure 1.1). Since strings may turn out to be extremely tiny, it may be difficult to observe directly the string-like nature of particles.

Are we sure that string theory is a good quantum theory of gravity? There is no complete certainty yet, but the evidence is very good. Indeed, the problems that occur when one tries to quantize Einstein's theory do not seem to appear in string theory.

For a theory as ambitious as string theory, a certain degree of uniqueness is clearly desirable. It would be somewhat disappointing to have several consistent candidates for a theory of all interactions. The first sign that string theory is rather unique is that it does not have adjustable dimensionless parameters. As we mentioned before, the Standard Model of particle physics has about twenty parameters that must be adjusted to some precise values. A theory with adjustable dimensionless parameters is not really unique. When the parameters are set to different values one obtains different theories with potentially different predictions. String theory has one dimensionful parameter, the string length ℓ_s. Its value can be roughly imagined as the typical size of strings.

Another intriguing sign of the uniqueness of string theory is the fact that the dimensionality of spacetime is fixed. Our physical spacetime is four-dimensional, with one time dimension and three space dimensions. In the Standard Model this information is used to build the theory, it is not derived. In string theory, on the other hand, the number of spacetime dimensions emerges from a calculation. The answer is not four, but rather ten.

Some of these dimensions may hide from plain view if they curl up into a space that is small enough to escape detection in experiments done with low energies. If string theory is correct, some mechanism must ensure that the observable dimensionality of spacetime is four.

The lack of adjustable dimensionless parameters is a sign of the uniqueness of string theory: it means that the theory cannot be deformed or changed continuously by changing these parameters. But there could be other theories that cannot be reached by continuous deformations. So how many string theories are there?

Let us begin by noting two broad subdivisions. There are open strings and there are closed strings. Open strings have two endpoints, while closed strings have no endpoints. One can consider theories with only closed strings and theories with both open and closed strings. Since open strings generally can close to form closed strings, we do not consider theories with only open strings. The second subdivision is between bosonic string theories and superstring theories. Bosonic strings live in 26 dimensions, and all of their vibrations represent bosons. Since they lack fermions, bosonic string theories are not realistic. They are, however, much simpler than the superstrings, and most of the important concepts in string theory can be explained in the context of bosonic strings. The superstrings live in ten-dimensional spacetime, and their spectrum of states includes bosons *and* fermions. In fact, these two sets of particles are related by supersymmetry. Supersymmetry is therefore an important ingredient in string theory. All realistic models of string theory are built from superstrings. In all string theories the graviton appears as a vibrational mode of closed strings. In string theory gravity is unavoidable.

By the mid 1980s five ten-dimensional superstring theories were known to exist. In the years that followed, many interrelations between these theories were found. Moreover, another theory was discovered by taking a certain strong coupling limit of one of the superstrings. This theory is eleven-dimensional and has been dubbed M-theory, for lack of a better name. It has now become clear that the five superstrings and M-theory are only facets or different limits of a *single* unique theory! At present, this unique theory remains fairly mysterious. It is not yet clear whether or not the set of bosonic string theories is connected to the web of superstring theories.

All in all, we see that string theory is a truly unified and possibly unique theory. It is a candidate for a unified theory of physics, a theory Albert Einstein tried to find ever since his discovery of general relativity. Einstein would have been surprised, or perhaps disturbed, by the prominent role that quantum mechanics plays in string theory. But string theory appears to be a worthy successor of general relativity. It is almost certain that string theory will give rise to a new conception of spacetime. The prominence of quantum mechanics in string theory would not have surprised Paul Dirac. His writings on quantization suggest that he felt that deep quantum theories arise from the quantization of classical physics. This is precisely what happens in string theory. This book will explain in detail how string theory,

at least in its simplest form, is nothing but the quantum mechanics of classical relativistic strings.

1.3 String theory and its verification

It should be said at the outset that, as of yet, there has been no experimental verification of string theory. In order to have experimental verification one needs a sharp prediction. It has been difficult to obtain such a prediction. String theory is still at an early stage of development, and it is not so easy to make predictions with a theory that is not well understood. Still, some interesting possibilities have emerged.

As we mentioned earlier, superstring theory requires a ten-dimensional spacetime: one dimension of time and nine of space. If string theory is correct, extra spatial dimensions must exist, even if we have not seen them yet. Can we test the existence of these extra dimensions? If the extra dimensions are the size of the Planck length ℓ_{P} (the length scale associated with four-dimensional gravity), they will remain beyond direct detection, perhaps forever. Indeed, $\ell_{\mathrm{P}} \sim 10^{-33}$ cm, and this distance is many orders of magnitude smaller than 10^{-16} cm, which is roughly the smallest distance that has been explored with particle accelerators. This scenario was deemed to be most likely. It was assumed that in string theory the length scale ℓ_s coincides with the Planck length, in which case extra dimensions would be of Planck length, as well.

It turns out, however, that string theory allows extra dimensions that are as large as a tenth of a millimeter! Surprisingly, extra dimensions that large may have gone undetected. To make this work out, the string length ℓ_s is taken to be of the order of 10^{-18} cm. Moreover, our three-dimensional space emerges as a hypersurface embedded inside the nine-dimensional space. The hypersurface, or higher-dimensional membrane, is called a D-brane. D-branes are real, physical objects in string theory. In this setup, the presence of large extra dimensions is tested by gravitational experiments. Extra dimensions much larger than ℓ_{P} but still very small may be detected with particle accelerators. If extra dimensions are detected, this would be strong evidence for string theory. We discuss the subject of large extra dimensions in Chapter 3.

A striking confirmation of string theory may result from the discovery of a cosmic string. Left-over from early universe processes, a cosmic string can stretch across the observable universe and may be detected via gravitational lensing or, more indirectly, through the detection of gravitational waves. No cosmic strings have been detected to date, but the searches have not been exhaustive and they continue. If found, a cosmic string must be studied in detail to confirm that it is a string from string theory and not the kind of string that can arise from conventional theories of particle physics. We discuss the subject of cosmic strings in Chapter 7.

Another interesting possibility has to do with supersymmetry. If we start with a ten-dimensional superstring theory and compactify the six extra dimensions, the resulting four-dimensional theory is, in many cases, supersymmetric. No unique predictions have emerged for the specific details of the four-dimensional theory, but supersymmetry

may be a rather generic feature. An experimental discovery of supersymmetry in future accelerators would suggest very strongly that string theory is on the right track.

Leaving aside predictions of new phenomena, we must ask whether the Standard Model emerges from string theory. It should, since string theory is supposed to be a unified theory of all interactions, and it must therefore reduce to the Standard Model for sufficiently low energies. While string theory certainly has room to include all known particles and interactions, and this is very good news indeed, no one has yet been able to show that they actually emerge in fine detail. In Chapter 21 we will study some models which use D-branes and have an uncanny resemblance to the world as we know it. In these models the particle content is in fact *precisely* that of the Standard Model (the particles are obtained with zero mass, however, and it is not clear whether the process that gives them mass can work out correctly). Our four-dimensional world is part of the D-branes, but these D-branes happen to have more than three spatial directions. The additional D-brane dimensions are wrapped on the compact space (we will learn how to imagine such configurations!). The gauge bosons and the matter particles in the model arise from vibrations of open strings that stretch between D-branes. As we will learn, the endpoints of open strings must remain attached to the D-branes. If you wish, the musical analogy for strings is improved. Just as the strings of a violin are held stretched by pegs, the D-branes hold fixed the endpoints of the open strings whose lowest vibrational modes could represent the particles of the Standard Model!

String theory shares with Einstein's gravity a problematic feature. Einstein's equations of gravitation admit many cosmological solutions. Each solution represents a consistent universe, but only one of them represents *our* observable universe. It is not easy to explain what selects the physical solution, but in cosmology this is done using arguments based on initial conditions, symmetry, and simplicity. The smaller the number of solutions a theory has, the more predictive it is. If the set of solutions is characterized by continuous parameters, selecting a solution is equivalent to adjusting the values of the parameters. In this way, a theory whose formulation requires no adjustable parameters may generate adjustable parameters through its solutions! It seems clear that in string theory the set of solutions (string models) is characterized by both discrete and continuous parameters.

In order to reproduce the Standard Model it seems clear that the string model must not have continuous parameters; such parameters imply the existence of massless fields that have not been observed. It was not easy to find models without continuous parameters, but that became possible recently in the context of flux compactifications; models in which the extra dimensions are threaded by analogs of electric and magnetic fields. There is an extraordinary large number of such models, certainly more than 10^{500} of them. There may be even more models that manage to avoid continuous parameters by other means. Physicists speak of a vast *landscape* of string solutions or models.

In this light we can wonder what are the possible outcomes of the search for a realistic string model. One possible outcome (the worst one) is that no string model in the landscape reproduces the Standard Model. This would rule out string theory. Another possible outcome (the best one) is that one string model reproduces the Standard Model. Moreover,

the model represents a well-isolated point in the landscape. The parameters of the Standard Model are thus predicted. The landscape may be so large that a strange possibility emerges: many string models with almost identical properties all of which are consistent with the Standard Model to the accuracy that it is presently known. In this possibility there is a loss of predictive power. Other outcomes may be possible.

String theorists sometimes say that string theory has already made at least one successful *prediction*: it predicted gravity! (I heard this from John Schwarz.) There is a bit of jest in saying so – after all, gravity is the oldest known force in nature. I believe, however, that there is a very substantial point to be made here. String theory is the quantum mechanics of a relativistic string. In no sense whatsoever is gravity put into string theory by hand. It is a complete surprise that gravity emerges in string theory. Indeed, none of the vibrations of the *classical* relativistic string correspond to the particle of gravity. It is a truly remarkable fact that we find the particle of gravity among the *quantum* vibrations of the relativistic string. You will see in detail how this happens as you progress through this book. The striking quantum emergence of gravitation in string theory has the full flavor of a prediction.

1.4 Developments and outlook

String theory has been a very stimulating and active area of research ever since Michael Green and John Schwarz showed in 1984 that superstrings are not afflicted with fatal inconsistencies that threaten similar particle theories in ten dimensions. Much progress has been made since then.

String theory has provided new and powerful tools for the understanding of conventional particle physics theories, gauge theories in particular. These are the kinds of theories that are used to formulate the Standard Model. Close cousins of these gauge theories arise on string theory D-branes. We examine D-branes and the theories that arise on them in detail beginning in Chapter 15. A remarkable physical equivalence between a certain four-dimensional gauge theory and a closed superstring theory (the AdS/CFT correspondence) is discussed in Chapter 23. As we will explain, the correspondence has been used to understand hydrodynamical properties of the quark-gluon plasma created in the collision of gold nuclei at heavy ion colliders.

String theory has also made good strides towards a statistical mechanics interpretation of black hole entropy. We know from the pioneering work of Jacob Bekenstein and Stephen Hawking that black holes have both entropy and temperature. In statistical mechanics these properties arise if a system can be constructed in many degenerate ways using its basic constituents. Such an interpretation is not available in Einstein's gravitation, where black holes seem to have few, if any, constituents. In string theory, however, certain black holes can be built by assembling together various types of D-branes and strings in a controlled manner. For such black holes, the predicted Bekenstein entropy is obtained by counting the ways in which they can be built with their constituent D-branes and strings. In fact, the

class of black holes amenable to string theory analysis continues to grow. We discuss this important development in Chapter 22.

String theory will be needed to study cosmology of the Very Early Universe. String theory may provide a concrete model for the realization of inflation – a period of dramatic exponential expansion that the universe is likely to have experienced at the earliest times. The theory of inflation suggests that our universe is a growing bubble or region inside a space that continues to inflate for eternity. Bubbles continue to emerge forever and some have remarked that every model in the landscape may be physically realized in some bubble. Inflation does not appear to be eternal in the past, so some kind of beginning seems necessary. The deepest mysteries of the universe seem to lie hidden in a regime where classical general relativity surely breaks down. String theory should allow us to peer into this unknown realm. Some day we may be able to understand how the universe comes into being, if it does, or how the universe could have existed forever in the past, if it did.

Most likely, answering such questions will require a mastery of string theory that goes beyond our present abilities. String theory is in fact an unfinished theory. Much has been learned about it, but in reality we have no complete formulation of the theory. A comparison with Einstein's theory is illuminating. Einstein's equations for general relativity are elegant and geometrical. They embody the conceptual foundation of the theory and feel completely up to the task of describing gravitation. No similar equations are known for string theory, and the conceptual foundation of the theory remains largely unknown. String theory is an exciting research area because the central ideas remain to be found.

Describing nature and formulating the theory – those remain the present-day challenges of string theory. If surmounted, we will have a theory of all interactions, allowing us to understand the fate of spacetime and the mysteries of a quantum mechanical universe. With such high stakes, physicists are likely to investigate string theory until definite answers are found.

2 Special relativity and extra dimensions

The word relativistic, as used in the term "relativistic strings," indicates consistency with Einstein's theory of special relativity. We review special relativity and introduce the light-cone frame, light-cone coordinates, and light-cone energy. We then turn to the idea of additional, compact space dimensions and show with an example from quantum mechanics that, if small, these dimensions have little effect at low energies.

2.1 Units and parameters

Units are nothing other than fixed quantities that we use for purposes of reference. A measurement involves finding the unit-free ratio of an observable quantity to the appropriate unit. Consider, for example, the definition of a second in the international system of units (SI system). The SI second (s) is defined to be the duration of 9 192 631 770 periods of the radiation emitted in the transition between the two hyperfine levels of the cesium-133 atom. When we measure the time elapsed between two events, we are really counting a unit-free, or dimensionless, number: the number that tells us how many seconds fit between the two events or, alternatively, how many periods of the cesium radiation fit between the two events. The same goes for length. The unit called the meter (m) is nowadays defined as the distance traveled by light in a certain fraction of a second (1/299 792 458 of a second, to be precise). Mass introduces a third unit, the prototype kilogram (kg), kept safely in Sèvres, France.

When doing dimensional analysis, we denote the units of length, time, and mass by L, T, and M, respectively. These are called the three basic units. A force, for example, has units

$$[F] = MLT^{-2}\,, \tag{2.1}$$

where $[X]$ denotes the units of the quantity X. Equation (2.1) follows from Newton's law that equates the force on an object to the product of its mass and its acceleration. The newton (N) is the SI unit of force, and it equals kg·m/s^2.

It is interesting that no additional basic units are needed to describe other quantities. Consider, for example, electric charge. Do we need a new unit to describe charge? Not

really. This is easy to see in Gaussian units. In these units, Coulomb's law for the force $|\vec{F}|$ between two charges q_1 and q_2 separated by a distance r reads

$$|\vec{F}| = \frac{|q_1 q_2|}{r^2}. \tag{2.2}$$

The units of charge are fixed in terms of other units because we have a force law where charges appear and all other quantities have known units. The esu is the Gaussian unit of charge, and it is defined by stating that two charges of one esu each, placed at a distance of one centimeter apart, repel each other with a force of one dyne (the Gaussian unit of force, 10^{-5} N). Thus

$$\text{esu}^2 = \text{dyne} \cdot \text{cm}^2 = 10^{-5}\ \text{N} \cdot (10^{-2}\text{m})^2 = 10^{-9}\ \text{N} \cdot \text{m}^2. \tag{2.3}$$

It follows from this equation that

$$[\text{esu}^2] = [\,\text{N} \cdot \text{m}^2\,], \tag{2.4}$$

and, using (2.1), finally we get

$$[\text{esu}] = M^{1/2} L^{3/2}\, T^{-1}. \tag{2.5}$$

This expresses the esu in terms of the three basic units.

In SI units, charge is measured in coulombs (C). The situation in SI units is a little more intricate, but the essential point is the same. A coulomb is defined in SI units as the amount of charge carried by a current of one ampere (A) in one second. The ampere itself is defined as the amount of current that, when carried by two wires separated by a distance of one meter, produces a force of 2×10^{-7} N/m. The coulomb, as opposed to the esu, is not expressed in terms of meters, kilograms, and seconds. Coulomb's law in SI units is

$$|\vec{F}| = \frac{1}{4\pi\epsilon_0}\frac{|q_1 q_2|}{r^2}, \quad \text{with} \quad \frac{1}{4\pi\epsilon_0} = 8.99 \times 10^9\ \frac{\text{N} \cdot \text{m}^2}{\text{C}^2}. \tag{2.6}$$

Note the presence of C^{-2} in the definition of the constant prefactor. Since each charge carries one factor of C, all the factors of C cancel in the calculation of the force. Two charges of one coulomb each, placed one meter apart, will each experience a force of 8.99×10^9 N. This fact allows you to deduce (Problem 2.1) how many esus there are in a coulomb. Even though we do not write coulombs in terms of other units, this is just a matter of convenience. Coulombs and esus are related, and esus are written in terms of the three basic units.

When we speak of parameters in a theory, it is convenient to distinguish between dimensionful parameters and dimensionless parameters. Consider, for example, a theory in which there are three types of particles with masses m_1, m_2, and m_3. We can think of the theory as having one dimensionful parameter, the mass m_1 of the first particle, say, and two dimensionless parameters, the mass ratios m_2/m_1 and m_3/m_1.

String theory is said to have no adjustable parameters. By this it is meant that no dimensionless parameter is needed to formulate string theory. String theory does, however, have one dimensionful parameter. That parameter is the string length ℓ_s. This length sets the

scale in which the theory operates. In the early 1970s, when string theory was first being formulated, the theory was thought to be a theory of hadrons. Back then, the string length was taken to be comparable to the nuclear scale. Nowadays, we think that string theory is a theory of fundamental forces and interactions. Accordingly, we set the string length to be much smaller than the nuclear scale.

2.2 Intervals and Lorentz transformations

Special relativity is based on the experimental fact that the speed of light ($c \simeq 3 \times 10^8$ m/s) is the same for all inertial observers. This fact leads to some rather surprising conclusions. Newtonian intuition about the absolute nature of time, the concept of simultaneity, and other familiar ideas must be revised. In comparing the coordinates of events, two inertial observers, henceforth called Lorentz observers, find that the appropriate coordinate transformations mix space and time.

In special relativity, events are characterized by the values of four coordinates: a time coordinate t and three spatial coordinates x, y, and z. It is convenient to collect these four numbers in the form (ct, x, y, z), where the time coordinate is scaled by the speed of light so that all coordinates have units of length. To make the notation more uniform, we use indices to relabel the space and time coordinates as follows:

$$x^\mu = (x^0, x^1, x^2, x^3) \equiv (ct, x, y, z). \tag{2.7}$$

Here the superscript μ takes the four values $0, 1, 2,$ and 3. The x^μ are spacetime coordinates.

Consider a Lorentz frame S in which two events are represented by the coordinates x^μ and $x^\mu + \Delta x^\mu$. Consider now a second Lorentz frame S', in which the same two events are described by the coordinates x'^μ and $x'^\mu + \Delta x'^\mu$, respectively. In general, not only are the coordinates x^μ and x'^μ different, so too are the coordinate differences Δx^μ and $\Delta x'^\mu$. On the other hand, both observers will agree on the value of the invariant interval Δs^2. This interval is defined by

$$-\Delta s^2 \equiv -(\Delta x^0)^2 + (\Delta x^1)^2 + (\Delta x^2)^2 + (\Delta x^3)^2. \tag{2.8}$$

Note the minus sign in front of $(\Delta x^0)^2$, as opposed to the plus sign appearing before the spacelike differences $(\Delta x^i)^2$ $(i = 1, 2, 3)$. This sign encodes the fundamental difference between time and space coordinates. The agreement on the value of the intervals is expressed as

$$-(\Delta x^0)^2 + (\Delta x^1)^2 + (\Delta x^2)^2 + (\Delta x^3)^2 = -(\Delta x'^0)^2 + (\Delta x'^1)^2 + (\Delta x'^2)^2 + (\Delta x'^3)^2, \tag{2.9}$$

or, in brief:

$$\Delta s^2 = \Delta s'^2. \tag{2.10}$$

The minus sign on the left-hand side of (2.8) implies that $\Delta s^2 > 0$ for events that are *timelike separated*. Timelike separated events are events for which

$$(\Delta x^0)^2 > (\Delta x^1)^2 + (\Delta x^2)^2 + (\Delta x^3)^2. \tag{2.11}$$

The history of a particle is represented in spacetime as a curve, the *world-line* of the particle. Any two events on the world-line of a particle are timelike separated, because no particle can move faster than light and therefore the distance light would have traveled in the time interval that separates the events must be larger than the space separation between the events. This is the content of (2.11). You at the time you were born and you at this moment are timelike separated events: a long time has passed and you have not gone that far. Events connected by the world-line of a photon are said to be *lightlike separated*. For such a pair of events, we have $\Delta s^2 = 0$, because in this case the two sides of (2.11) are identical: the spatial separation between the events coincides with the distance that light would have traveled in the time that separates the events. Two events for which $\Delta s^2 < 0$ are said to be *spacelike separated*. Events that are simultaneous in a Lorentz frame but occur at different positions in that same frame are spacelike separated. It is because Δs^2 can be negative that it is not written as $(\Delta s)^2$. For timelike separated events, however, we define

$$\Delta s \equiv \sqrt{\Delta s^2} \quad \text{if} \quad \Delta s^2 > 0 \quad \text{(timelike interval)}. \tag{2.12}$$

Many times it is useful to consider events that are infinitesimally close to each other. Small coordinate differences are needed to define velocities and are also useful in general relativity. Infinitesimal coordinate differences are written as dx^μ, and the associated invariant interval is written as ds^2. Following (2.8), we have

$$-ds^2 = -(dx^0)^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2. \tag{2.13}$$

The equality of intervals is the statement

$$ds^2 = ds'^2. \tag{2.14}$$

A very useful notation can be motivated by trying to simplify the expression for the invariant ds^2. To do this, we introduce symbols that carry subscripts instead of superscripts. Let us define

$$dx_0 \equiv -dx^0, \quad dx_1 \equiv dx^1, \quad dx_2 \equiv dx^2, \quad dx_3 \equiv dx^3. \tag{2.15}$$

The only significant change is the inclusion of a minus sign for the zeroth component. All together, we write

$$dx_\mu = (dx_0, dx_1, dx_2, dx_3) \equiv (-dx^0, dx^1, dx^2, dx^3). \tag{2.16}$$

Now we can rewrite ds^2 in terms of dx^μ and dx_μ:

$$\begin{aligned} -ds^2 &= -(dx^0)^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2 \\ &= dx_0 dx^0 + dx_1 dx^1 + dx_2 dx^2 + dx_3 dx^3, \end{aligned} \tag{2.17}$$

and we see that the minus sign in (2.13) is gone. The invariant interval has become

$$-ds^2 = \sum_{\mu=0}^{3} dx_\mu dx^\mu. \tag{2.18}$$

Throughout the rest of the text we will use Einstein's summation convention. In this convention, indices repeated in a single term are to be summed over the appropriate set of values. We do not consider indices to be repeated when they appear in different terms. For example, there are no sums implied by $a^\mu + b^\mu$ or $a^\mu = b^\mu$, but there is an implicit sum in $a^\mu b_\mu$. A repeated index must appear once as a subscript and once as a superscript and should not appear more than twice in any one term. The letter chosen for the repeated index is not important; thus $a^\mu b_\mu$ is the same as $a^\nu b_\nu$. Because of this, repeated indices are sometimes called dummy indices! Using the summation convention, we can rewrite (2.18) as

$$-ds^2 = dx_\mu dx^\mu. \tag{2.19}$$

Just as we did for finite coordinate differences in (2.12), for infinitesimal timelike intervals we define the quantity

$$ds \equiv \sqrt{ds^2} \quad \text{if} \quad ds^2 > 0 \quad \text{(timelike interval)}\,. \tag{2.20}$$

We can also express the interval ds^2 using the Minkowski metric $\eta_{\mu\nu}$. This is done by writing

$$-ds^2 = \eta_{\mu\nu}\, dx^\mu dx^\nu\,. \tag{2.21}$$

Equation (2.21), by itself, does not determine the metric $\eta_{\mu\nu}$. If, in addition, we require $\eta_{\mu\nu}$ to be symmetric under an exchange in the order of its indices,

$$\eta_{\mu\nu} = \eta_{\nu\mu}\,, \tag{2.22}$$

then this together with (2.21) completely determines a metric called the Minkowski metric. It is reasonable to declare that $\eta_{\mu\nu}$ be symmetric for, as we shall now see, any antisymmetric component would be irrelevant.

Any two-index object $M_{\mu\nu}$ can be decomposed into a symmetric part and an antisymmetric part:

$$M_{\mu\nu} = \frac{1}{2}(M_{\mu\nu} + M_{\nu\mu}) + \frac{1}{2}(M_{\mu\nu} - M_{\nu\mu})\,. \tag{2.23}$$

The first term on the right-hand side, the symmetric part of M, is invariant under exchange of the indices μ and ν. The second term on the right-hand side, the antisymmetric part of M, changes sign under exchange of the indices μ and ν. If $\eta_{\mu\nu}$ had an antisymmetric part $\xi_{\mu\nu}$ $(= -\xi_{\nu\mu})$, then its contribution would drop out of the right-hand side of (2.21). We can see this as follows:

$$\xi_{\mu\nu}\, dx^\mu dx^\nu = (-\xi_{\nu\mu})\, dx^\mu dx^\nu = -\xi_{\mu\nu}\, dx^\nu dx^\mu = -\xi_{\mu\nu}\, dx^\mu dx^\nu. \tag{2.24}$$

In the first step, we used the antisymmetry of $\xi_{\mu\nu}$. In the second step, we relabeled the dummy indices: the μ were changed into ν and vice versa. In the third step, we switched

the order of the dx^μ and dx^ν factors. The result is that $\xi_{\mu\nu}dx^\mu dx^\nu$ is identical to minus itself, and therefore it vanishes. It would be useless for $\eta_{\mu\nu}$ to have an antisymmetric part, so we simply declare that it has none.

Since repeated indices are summed over, equation (2.21) means that

$$-ds^2 = \eta_{00}dx^0dx^0 + \eta_{01}dx^0dx^1 + \eta_{10}dx^1dx^0 + \eta_{11}dx^1dx^1 + \cdots. \tag{2.25}$$

Comparing with (2.17) and using (2.22), we see that $\eta_{00} = -1$, $\eta_{11} = \eta_{22} = \eta_{33} = 1$, and all other components vanish. We collect these values in matrix form:

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2.26}$$

In this equation, which follows a common identification of two-index objects with matrices, we think of μ, the first index in η, as the row index, and ν, the second index in η, as the column index. The Minkowski metric can be used to "lower indices." Indeed, equation (2.15) can be rewritten as

$$dx_\mu = \eta_{\mu\nu}\, dx^\nu. \tag{2.27}$$

If we are handed a set of quantities b^μ, we always define

$$b_\mu \equiv \eta_{\mu\nu}b^\nu. \tag{2.28}$$

Given objects a^μ and b^μ, the relativistic *scalar product* $a \cdot b$ is defined as

$$a \cdot b \equiv a^\mu b_\mu = \eta_{\mu\nu}\, a^\mu b^\nu = -a^0b^0 + a^1b^1 + a^2b^2 + a^3b^3. \tag{2.29}$$

Applied to (2.19), this gives $-ds^2 = dx \cdot dx$. Note that $a^\mu b_\mu = a_\mu b^\mu$.

It is convenient to introduce the matrix inverse for $\eta_{\mu\nu}$. Written conventionally as $\eta^{\mu\nu}$, it is given by

$$\eta^{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2.30}$$

You can see by inspection that this matrix is indeed the inverse of the matrix in (2.26). When thinking of $\eta^{\mu\nu}$ as a matrix, as with $\eta_{\mu\nu}$, the first index is a row index and the second index is a column index. In index notation the inverse property is written as

$$\eta^{\nu\rho}\,\eta_{\rho\mu} = \delta^\nu_\mu\,, \tag{2.31}$$

where the Kronecker delta δ^ν_μ is defined by

$$\delta^\nu_\mu = \begin{cases} 1 & \text{if } \mu = \nu \\ 0 & \text{if } \mu \neq \nu. \end{cases} \tag{2.32}$$

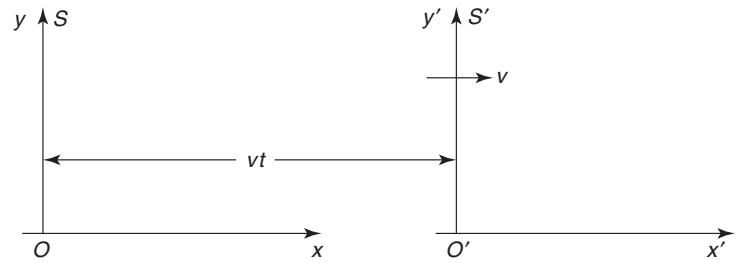

Fig. 2.1 **Two Lorentz frames connected by a boost. S' is boosted along the $+x$ direction of S with boost parameter $\beta = v/c$.**

Note that the repeated index ρ in (2.31) produces the desired matrix multiplication. The Kronecker delta can be thought of as the index representation of the identity matrix. The metric with upper indices can be used to "raise indices." Using (2.28) and (2.31), we get

$$\eta^{\rho\mu}\, b_\mu = \eta^{\rho\mu}\ (\eta_{\mu\nu}\, b^\nu) = (\eta^{\rho\mu}\, \eta_{\mu\nu})\, b^\nu = \delta^\rho_\nu\, b^\nu = b^\rho. \tag{2.33}$$

The lower μ index of b_μ was raised by $\eta^{\rho\mu}$ to become an upper ρ index. The last step in the above calculation needs a little explanation: $\delta^\rho_\nu\, b^\nu = b^\rho$ because as we sum over ν, δ^ρ_ν vanishes unless $\nu = \rho$, in which case it equals one.

Lorentz transformations are the relations between coordinates in two different inertial frames. Consider a frame S and a frame S', that is moving along the positive x direction of the S frame with a velocity v, as shown in Figure 2.1. Assume the coordinate axes for both systems are parallel, and that the origins coincide at the common time $t = t' = 0$. We say that S' is boosted along the x direction with velocity parameter $\beta \equiv v/c$. The Lorentz transformations in this case read

$$\begin{aligned} x' &= \gamma(x - \beta ct), \\ ct' &= \gamma(ct - \beta x), \\ y' &= y, \\ z' &= z, \end{aligned} \tag{2.34}$$

where the Lorentz factor γ is given by

$$\gamma \equiv \frac{1}{\sqrt{1-\beta^2}} = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}\,. \tag{2.35}$$

Using indices, and changing the order of the first two equations, we arrive at

$$\begin{aligned} x'^0 &= \gamma(x^0 - \beta x^1), \\ x'^1 &= \gamma(-\beta x^0 + x^1), \\ x'^2 &= x^2, \\ x'^3 &= x^3. \end{aligned} \tag{2.36}$$

In the above transformations, the coordinates x^2 and x^3 remain unchanged. These are the coordinates orthogonal to the direction of the boost. The inverse Lorentz transformations give the values of the x coordinates in terms of the x' coordinates. They are readily found by solving for the x in the above equations. The result is the same set of transformations with x and x' exchanged and with β replaced by $(-\beta)$, as required by symmetry.

The coordinates in the above equations satisfy the relation

$$(x^0)^2 - (x^1)^2 - (x^2)^2 - (x^3)^2 = (x'^0)^2 - (x'^1)^2 - (x'^2)^2 - (x'^3)^2 \,, \tag{2.37}$$

as you can show by direct computation. This is just the statement of invariance of the interval Δs^2 between two events: the first event is represented by $(0, 0, 0, 0)$ in both S and S', and the second event is represented by coordinates x^μ in S and x'^μ in S'. By definition, *Lorentz transformations are the linear transformations of coordinates that respect the equality* (2.37).

In general, we write a Lorentz transformation as the linear relation

$$x'^\mu = L^\mu{}_\nu \, x^\nu \,, \tag{2.38}$$

where the entries $L^\mu{}_\nu$ are constants that define the linear transformation. For the boost in (2.36), we have

$$[L] = L^\mu{}_\nu = \begin{pmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} . \tag{2.39}$$

In defining the matrix L as $[L] = L^\mu{}_\nu$, we are following the convention that the first index is a row index and the second index is a column index. This is why the lower index in $L^\mu{}_\nu$ is written to the right of the upper index.

The coefficients $L^\mu{}_\nu$ are constrained by equation (2.37). In index notation, this equation requires

$$\eta_{\alpha\beta} \, x^\alpha x^\beta = \eta_{\mu\nu} \, x'^\mu x'^\nu \,. \tag{2.40}$$

Using (2.38) twice on the right-hand side above gives

$$\eta_{\alpha\beta} \, x^\alpha x^\beta = \eta_{\mu\nu} \, (L^\mu{}_\alpha \, x^\alpha)(L^\nu{}_\beta \, x^\beta) = \eta_{\mu\nu} \, L^\mu{}_\alpha \, L^\nu{}_\beta \, x^\alpha \, x^\beta \,. \tag{2.41}$$

Equivalently, we have the equation

$$k_{\alpha\beta} \, x^\alpha \, x^\beta = 0 \,, \quad \text{with} \quad k_{\alpha\beta} \equiv \eta_{\mu\nu} \, L^\mu{}_\alpha \, L^\nu{}_\beta \, - \eta_{\alpha\beta} \,. \tag{2.42}$$

Since $k_{\alpha\beta} \, x^\alpha \, x^\beta = 0$ must hold for all values of the coordinates x, we find that

$$k_{\alpha\beta} + k_{\beta\alpha} = 0 \,, \tag{2.43}$$

as you should convince yourself by writing out the sums over α and β. Since $k_{\alpha\beta}$ is in fact symmetric under the exchange in the order of its indices, (2.43) implies $k_{\alpha\beta} = 0$. This means that

$$\eta_{\mu\nu} \, L^\mu{}_\alpha \, L^\nu{}_\beta = \eta_{\alpha\beta} \,. \tag{2.44}$$

Rewriting (2.44) to make it look more like matrix multiplication, we have

$$L^{\mu}{}_{\alpha}\ \eta_{\mu\nu}\ L^{\nu}{}_{\beta} = \eta_{\alpha\beta}\,. \tag{2.45}$$

The sum over the ν index works well: it is a column index in $\eta_{\mu\nu}$ and a row index in $L^{\nu}{}_{\beta}$. The μ index, however, is a row index in $L^{\mu}{}_{\alpha}$, while it should be a column index to match the row index in $\eta_{\mu\nu}$. Moreover, the α index in $L^{\mu}{}_{\alpha}$ is a column index, while it is a row index in $\eta_{\alpha\beta}$. This means that we should exchange the columns and rows of $L^{\mu}{}_{\alpha}$, which is the matrix operation of transposition. Therefore equation (2.45) can be rewritten as the matrix equation

$$L^{\mathrm{T}}\,\eta\,L = \eta\,. \tag{2.46}$$

Here η is the matrix whose entries are $\eta_{\mu\nu}$. This neat equation is the constraint that L must satisfy to be a Lorentz transformation.

An important property of Lorentz transformations can be deduced by taking the determinant of each side of equation (2.46). Since the determinant of a product is the product of the determinants, we get

$$(\det L^{\mathrm{T}})(\det\eta)(\det L) = \det\eta\,. \tag{2.47}$$

Cancelling the common factor of $\det\eta$ and recalling that the operation of transposition does not change a determinant, we find

$$(\det L)^2 = 1 \ \longrightarrow\ \det L = \pm 1\,. \tag{2.48}$$

You can check that $\det L = 1$ for the boost in (2.39). Since $\det L$ never vanishes, the matrix L is always invertible and, consequently, all Lorentz transformations are invertible linear transformations.

The set of Lorentz transformations includes boosts along each of the spatial coordinates. It also includes rotations of the spatial coordinates. Under a spatial rotation, the coordinates (x^0, x^1, x^2, x^3) of a point transform into coordinates (x'^0, x'^1, x'^2, x'^3), for which $x^0 = x'^0$, because time is unaffected. Since the spatial distance from a point to the origin is preserved under a rotation, we have

$$(x^1)^2 + (x^2)^2 + (x^3)^2 = (x'^1)^2 + (x'^2)^2 + (x'^3)^2\,. \tag{2.49}$$

This, together with $x^0 = x'^0$, implies that (2.37) holds. Therefore spatial rotations are Lorentz transformations.

Any set of four quantities which transforms under Lorentz transformations in the same way as the x^{μ} do is said to be a four-vector, or Lorentz vector. When we use index notation and write b^{μ}, we mean that b^{μ} is a four-vector. Taking differentials of the linear equations (2.36), we see that the linear transformations that relate x' to x also relate dx' to dx. Therefore the differentials dx^{μ} define a Lorentz vector. In the spirit of index notation, a quantity with no free indices must be invariant under Lorentz transformations. A quantity has no free indices if it carries no index or if it contains only repeated indices, such as $a^{\mu}b_{\mu}$.

A four-vector a^{μ} is said to be timelike if $a^2 = a\cdot a < 0$, spacelike if $a^2 > 0$, and null if $a^2 = 0$. Recalling our discussion below (2.11), we see that the coordinate differences

between timelike-separated events define a timelike vector. Similarly, the coordinate differences between spacelike-separated events define a spacelike vector, and the coordinate differences between lightlike-separated events define a null vector.

Quick calculation 2.1 Verify that the invariant ds^2 is indeed preserved under the Lorentz transformations (2.36).

Quick calculation 2.2 Consider two Lorentz vectors a^μ and b^μ. Write the Lorentz transformations $a^\mu \to a'^\mu$ and $b^\mu \to b'^\mu$ analogous to (2.36). Verify that $a^\mu b_\mu$ is invariant under these transformations.

2.3 Light-cone coordinates

We now discuss a coordinate system that will be extremely useful in our study of string theory, the light-cone coordinate system. The quantization of the relativistic string can be worked out most directly using light-cone coordinates. There is a different approach to the quantization of the relativistic string, in which no special coordinates are used. This approach, called Lorentz covariant quantization, is discussed briefly in Chapter 24. Lorentz quantization is very elegant, but a full discussion requires a great deal of background material. We will use light-cone coordinates to quantize strings in this book.

We define the two light-cone coordinates x^+ and x^- as two independent linear combinations of the time coordinate and a chosen spatial coordinate, conventionally taken to be x^1. This is done by writing

$$x^+ \equiv \frac{1}{\sqrt{2}}(x^0 + x^1),$$
$$x^- \equiv \frac{1}{\sqrt{2}}(x^0 - x^1). \tag{2.50}$$

The coordinates x^2 and x^3 play no role in this definition. In the light-cone coordinate system, (x^0, x^1) is traded for (x^+, x^-), but the other two coordinates x^2, x^3 are kept. Thus, the complete set of light-cone coordinates is (x^+, x^-, x^2, x^3).

The new coordinates x^+ and x^- are called light-cone coordinates because the associated coordinate axes are the world-lines of beams of light emitted from the origin along the x^1 axis. For a beam of light moving in the positive x^1 direction, we have $x^1 = ct = x^0$, and thus $x^- = 0$. The line $x^- = 0$ is, by definition, the x^+ axis (Figure 2.2). For a beam of light moving in the negative x^1 direction, we have $x^1 = -ct = -x^0$, and thus $x^+ = 0$. This corresponds to the x^- axis. The $x^\pm$ axes are lines at 45° with respect to the x^0, x^1 axes.

Can we think of x^+, or perhaps x^-, as a new time coordinate? Yes. In fact, both have equal right to be called a time coordinate, although neither one is a time coordinate in the standard sense of the word. Light-cone time is not quite the same as ordinary time.

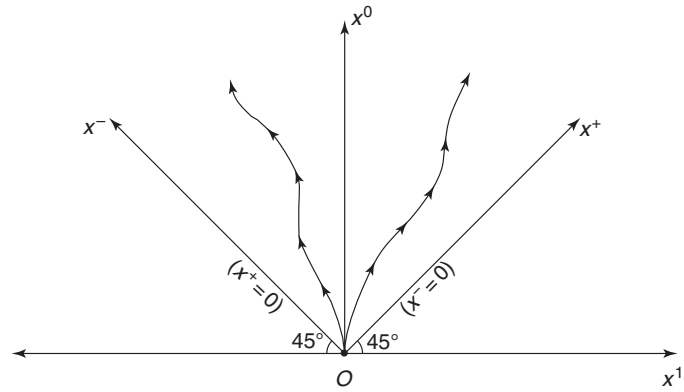

Fig. 2.2 **A spacetime diagram with x^1 and x^0 represented as orthogonal axes. Shown are the light-cone axes $x^{\pm} = 0$. The curves with arrows are possible world-lines of physical particles.**

Perhaps the most familiar property of time is that it goes forward for any physical motion of a particle. Physical motion starting at the origin is represented in Figure 2.2 as curves that remain within the light-cone and whose slopes never go below 45°. For all these curves, both x^+ and x^- increase as we follow the arrows. The only subtlety is that, for special light rays, light-cone time will freeze! As we saw above, x^+ remains constant for a light ray in the negative x^1 direction, while x^- remains constant for a light ray in the positive x^1 direction.

For definiteness, we will take x^+ to be the *light-cone time* coordinate. Accordingly, we will think of x^- as a spatial coordinate. Of course, these light-cone time and space coordinates will be somewhat strange.

Taking differentials of (2.50), we readily find that

$$2\,dx^+dx^- = (dx^0 + dx^1)\,(dx^0 - dx^1) = (dx^0)^2 - (dx^1)^2\,. \tag{2.51}$$

It follows that the invariant interval (2.13), expressed in terms of the light-cone coordinates (2.50), takes the form

$$-ds^2 = -2\,dx^+dx^- + (dx^2)^2 + (dx^3)^2\,. \tag{2.52}$$

The symmetry in the definitions of x^+ and x^- is evident here. Notice that, if we are given ds^2, solving for dx^- or for dx^+ does not require us to take a square root. This is a very important feature of light-cone coordinates, as we will see in Chapter 9.

How do we represent (2.52) with index notation? We still need indices that run over four values, but this time the values will be called

$$+\,,\,-\,,\,2\,,\,3. \tag{2.53}$$

Just as we did in (2.21), we write

$$-ds^2 = \hat{\eta}_{\mu\nu} dx^\mu dx^\nu . \tag{2.54}$$

Here we have introduced a light-cone metric $\hat{\eta}$ which, like the Minkowski metric, is also defined to be symmetric under the exchange of its indices. Expanding this equation, and comparing with (2.52), we find

$$\hat{\eta}_{+-} = \hat{\eta}_{-+} = -1 , \quad \hat{\eta}_{++} = \hat{\eta}_{--} = 0 . \tag{2.55}$$

In the $(+, -)$ subspace, the diagonal elements of the light-cone metric vanish, but the off-diagonal elements do not. We also find that $\hat{\eta}$ does not couple the $(+, -)$ subspace to the $(2, 3)$ subspace:

$$\hat{\eta}_{+I} = \hat{\eta}_{-I} = 0 , \quad I = 2, 3 . \tag{2.56}$$

The matrix representation of the light-cone metric is

$$\hat{\eta}_{\mu\nu} = \begin{pmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} . \tag{2.57}$$

The light-cone components of any Lorentz vector a^μ are defined in analogy with (2.50):

$$\begin{aligned} a^+ &\equiv \frac{1}{\sqrt{2}} (a^0 + a^1) , \\ a^- &\equiv \frac{1}{\sqrt{2}} (a^0 - a^1) . \end{aligned} \tag{2.58}$$

The scalar product between vectors, shown in (2.29), can be written using light-cone components. This time we have

$$a \cdot b = -a^- b^+ - a^+ b^- + a^2 b^2 + a^3 b^3 = \hat{\eta}_{\mu\nu} a^\mu b^\nu . \tag{2.59}$$

The last equality follows immediately from summing over the repeated indices and using (2.57). The first equality needs a small computation. In fact, it suffices to check that

$$-a^- b^+ - a^+ b^- = -a^0 b^0 + a^1 b^1 . \tag{2.60}$$

This is quickly done using (2.3) and the analogous equations for $b^\pm$. We can also introduce lower light-cone indices. Consider the expression $a \cdot b = a_\mu b^\mu$, and expand the sum over the index μ using the light-cone labels:

$$a \cdot b = a_+ b^+ + a_- b^- + a_2 b^2 + a_3 b^3 . \tag{2.61}$$

Comparing with (2.59), we find that

$$a_+ = -a^- , \quad a_- = -a^+ . \tag{2.62}$$

When we lower or raise the zeroth index in a Lorentz frame, we get an extra sign. In light-cone coordinates, the indices of the first two coordinates switch and we get an extra sign.

Since physics described using light-cone coordinates looks unusual, we must develop an intuition for it. To do this, we will look at an example where the calculations are simple but the results are surprising.

Consider a particle moving along the x^1 axis with speed parameter $\beta = v/c$. At time $t = 0$, the positions x^1, x^2, and x^3 are all zero. Motion is nicely represented when the positions are expressed in terms of time:

$$x^1(t) = vt = \beta x^0, \quad x^2(t) = x^3(t) = 0. \tag{2.63}$$

How does this look in light-cone coordinates? Since x^+ is time and $x^2 = x^3 = 0$, we must simply express x^- in terms of x^+. Using (2.63), we find

$$x^+ = \frac{x^0 + x^1}{\sqrt{2}} = \frac{1+\beta}{\sqrt{2}}\, x^0. \tag{2.64}$$

As a result,

$$x^- = \frac{x^0 - x^1}{\sqrt{2}} = \frac{(1-\beta)}{\sqrt{2}}\, x^0 = \frac{1-\beta}{1+\beta}\, x^+ . \tag{2.65}$$

Since it relates light-cone position to light-cone time, we identify the ratio

$$\frac{dx^-}{dx^+} = \frac{1-\beta}{1+\beta} \tag{2.66}$$

as the light-cone velocity. How strange is this light-cone velocity? For light moving to the right ($\beta = 1$) it equals zero. Indeed, light moving to the right has zero light-cone velocity because x^- does not change at all. This is shown as line 1 in Figure 2.3. Suppose you have a particle moving to the right with high conventional velocity, so that $\beta \simeq 1$ (line 2 in the figure). Its light-cone velocity is then very small. A long light-cone time must pass for this particle to move a little in the x^- direction. Perhaps more interestingly, a static particle in standard coordinates (line 3) is moving quite fast in light-cone coordinates. When $\beta = 0$ the particle has unit light-cone speed. This light-cone speed increases as β grows negative: the numerator in (2.66) is larger than one and increasing, while the denominator is smaller than one and decreasing. For $\beta = -1$ (line 5), the light-cone velocity is infinite! While this seems odd, there is no clash with relativity. Light-cone velocities are just unusual. The light-cone is a frame in which kinematics has a nonrelativistic flavor and infinite velocities are possible. Note that light-cone coordinates were introduced as a change of coordinates, not as a Lorentz transformation. There is no Lorentz transformation that takes the coordinates (x^0, x^1, x^2, x^3) into coordinates $(x'^0, x'^1, x'^2, x'^3) = (x^+, x^-, x^2, x^3)$.

Quick calculation 2.3 Convince yourself that the last statement above is correct.

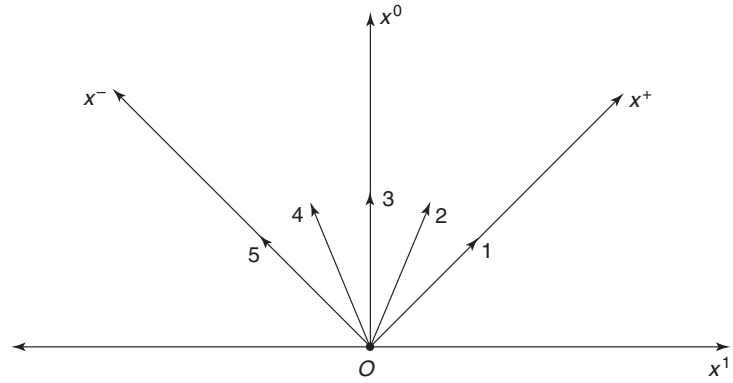

Fig. 2.3 World-lines of particles with various light-cone velocities. Particle 1 has zero light-cone velocity. The velocities increase through that of particle 5, which is infinite.

2.4 Relativistic energy and momentum

In special relativity there is a basic relationship between the rest mass m of a point particle, its relativistic energy E, and its relativistic momentum $\vec{p}$. This relationship is given by

$$\frac{E^2}{c^2} - \vec{p} \cdot \vec{p} = m^2 c^2. \tag{2.67}$$

The relativistic energy and momentum are given in terms of the rest mass and velocity by the following familiar relations:

$$E = \gamma m c^2, \quad \vec{p} = \gamma m \vec{v}. \tag{2.68}$$

Quick calculation 2.4 Verify that the above E and $\vec{p}$ satisfy (2.67).

Energy and momentum can be used to define a momentum four-vector, as we will prove shortly. This four-vector is

$$p^\mu = (p^0, p^1, p^2, p^3) \equiv \left(\frac{E}{c}, p_x, p_y, p_z\right). \tag{2.69}$$

Using the last two equations, we have

$$p^\mu = \left(\frac{E}{c}, \vec{p}\right) = m\gamma(c, \vec{v}). \tag{2.70}$$

We use (2.28) to lower the index in p^μ:

$$p_\mu = (p_0, p_1, p_2, p_3) = \eta_{\mu\nu} p^\nu = \left(-\frac{E}{c}, p_x, p_y, p_z\right). \tag{2.71}$$

The above expressions for p^μ and p_μ give

$$p^\mu p_\mu = -(p^0)^2 + \vec{p} \cdot \vec{p} = -\frac{E^2}{c^2} + \vec{p} \cdot \vec{p}\,, \tag{2.72}$$

and, making use of (2.67), we have

$$p^\mu p_\mu = -m^2c^2. \tag{2.73}$$

Since $p^\mu p_\mu$ has no free index it must be a Lorentz scalar. Indeed, all Lorentz observers agree on the value of the rest mass of a particle. Using the relativistic scalar product notation, condition (2.73) reads

$$p^2 \equiv p \cdot p = -m^2c^2. \tag{2.74}$$

A central concept in special relativity is that of *proper time*. Proper time is a Lorentz invariant measure of time. Consider a moving particle and two events along its trajectory. Different Lorentz observers record different values for the time interval between the two events. But now imagine that the moving particle is carrying a clock. The proper time elapsed is the time elapsed between the two events *on that clock*. By definition, it is an invariant: all observers of a particular clock must agree on the time elapsed on that clock!

Proper time enters naturally into the calculation of invariant intervals. Consider an invariant interval for the motion of a particle along the x axis:

$$-ds^2 = -c^2dt^2 + dx^2 = -c^2dt^2\,(1-\beta^2). \tag{2.75}$$

Now evaluate the interval using a Lorentz frame attached to the particle. This is a frame in which the particle does not move and time is recorded by the clock that is moving with the particle. In this frame $dx = 0$ and $dt = dt_p$ is the proper time elapsed. As a result,

$$-ds^2 = -c^2\,{dt_p}^2. \tag{2.76}$$

We cancel the minus signs and take the square root (using (2.20)) to find

$$ds = c\,dt_p. \tag{2.77}$$

This shows that, for timelike intervals, ds/c is the proper time interval. Similarly, cancelling minus signs and taking the square root of (2.75) gives

$$ds = cdt\sqrt{1-\beta^2} \;\longrightarrow\; \frac{dt}{ds} = \frac{\gamma}{c}. \tag{2.78}$$

Being a Lorentz invariant, ds can be used to construct new Lorentz vectors from old Lorentz vectors. For example, a velocity four-vector u^μ is obtained by taking the ratio of dx^μ and ds. Since dx^μ is a Lorentz vector and ds is a Lorentz scalar, the ratio is also a Lorentz vector:

$$u^\mu = c\frac{dx^\mu}{ds} = c\Big(\frac{d(ct)}{ds}, \frac{dx}{ds}, \frac{dy}{ds}, \frac{dz}{ds}\Big). \tag{2.79}$$

The factor of c is included to give u^μ the units of velocity. The components of u^μ can be simplified using the chain rule and (2.78). For example,

$$\frac{dx}{ds} = \frac{dx}{dt}\frac{dt}{ds} = \frac{v_x\gamma}{c}. \tag{2.80}$$

Back in (2.79), we find

$$u^\mu = \gamma(c, v_x, v_y, v_z) = \gamma(c, \vec{v}). \tag{2.81}$$

Comparing with (2.70), we see that the momentum four-vector is just mass times the velocity four-vector:

$$p^\mu = m u^\mu. \tag{2.82}$$

This confirms our earlier assertion that the components of p^μ form a four-vector. Since any four-vector transforms under Lorentz transformations as the x^μ do, we can use (2.36) to find that under a boost in the x-direction the p^μ transform as

$$\begin{aligned} \frac{E'}{c} &= \gamma\left(\frac{E}{c} - \beta\, p_x\right), \\ p'_x &= \gamma\left(-\beta\frac{E}{c} + p_x\right). \end{aligned} \tag{2.83}$$

2.5 Light-cone energy and momentum

The light-cone components p^+ and p^- of the momentum Lorentz vector are obtained using the rule (2.3):

$$\begin{aligned} p^+ &= \frac{1}{\sqrt{2}}(p^0 + p^1) = -p_-, \\ p^- &= \frac{1}{\sqrt{2}}(p^0 - p^1) = -p_+. \end{aligned} \tag{2.84}$$

Which component should be identified with light-cone energy? The naive answer would be p^+. In any Lorentz frame, both the time and energy are the zeroth components of their respective four-vectors. Since light-cone time was chosen to be x^+, we might conclude that light-cone energy should be taken to be p^+. This is not appropriate, however. Light-cone coordinates do not transform as Lorentz ones do, so we should be careful and examine this question in detail. Both $p^\pm$ are energy-like, since both are positive for physical particles. Indeed, from (2.67), and with $m \neq 0$, we have

$$p^0 = \frac{E}{c} = \sqrt{\vec{p}\cdot\vec{p} + m^2c^2} > |\,\vec{p}\,| \geq |p^1|\,. \tag{2.85}$$

As a result, $p^0 \pm p^1 > 0$, and thus $p^\pm > 0$. While both are plausible candidates for energy, the physically motivated choice turns out to be $-p_+$, which happens to coincide with p^-.

Before we explain this choice, let us first evaluate $p_\mu x^\mu$. In standard coordinates,

$$p \cdot x = p_0 x^0 + p_1 x^1 + p_2 x^2 + p_3 x^3. \tag{2.86}$$

In light-cone coordinates, using (2.61),

$$p \cdot x = p_+ x^+ + p_- x^- + p_2 x^2 + p_3 x^3. \tag{2.87}$$

In standard coordinates, $p_0 = -E/c$ appears together with the time x^0. In light-cone coordinates, p_+ appears together with the light-cone time x^+. We would therefore expect p_+ to be minus the light-cone energy.

Why is this pairing significant? Energy and time are conjugate variables. As you learned in quantum mechanics, the Hamiltonian operator measures energy and generates time evolution. The wavefunction of a point particle with energy E and momentum $\vec{p}$ is given by

$$\psi(t, \vec{x}) = \exp\Big(-\frac{i}{\hbar}(Et - \vec{p}\cdot\vec{x})\Big). \tag{2.88}$$

Indeed, this wavefunction satisfies the Schrödinger equation

$$i\hbar\,\frac{\partial\psi}{\partial x^0} = \frac{E}{c}\,\psi. \tag{2.89}$$

Similarly, light-cone time evolution and light-cone energy $E_{\rm lc}$ should be related by

$$i\hbar\,\frac{\partial\psi}{\partial x^+} = \frac{E_{\rm lc}}{c}\,\psi. \tag{2.90}$$

To find the x^+ dependence of the wavefunction, we recognize that

$$\psi(t, \vec{x}) = \exp\Big(\frac{i}{\hbar}\,(p_0 x^0 + \vec{p}\cdot\vec{x})\Big) = \exp\Big(\frac{i}{\hbar}\;p\cdot x\Big)\,, \tag{2.91}$$

and, using (2.87), we have

$$\psi(x) = \exp\Big(\frac{i}{\hbar}\,(p_+x^+ + p_-x^- + p_2x^2 + p_3x^3)\Big). \tag{2.92}$$

We can now return to (2.90) and evaluate:

$$i\hbar\frac{\partial\psi}{\partial x^+} = -p_+\psi \;\longrightarrow\; -p_+ = \frac{E_{\rm lc}}{c}. \tag{2.93}$$

This confirms our identification of $(-p_+)$ with light-cone energy. Since, presently, $-p_+ = p^-$, it is convenient to use p^- as the light-cone energy in order to eliminate the sign in the above equation:

$$p^- = \frac{E_{\rm lc}}{c}. \tag{2.94}$$

Some physicists like to raise and lower $+$ and $-$ indices to simplify expressions involving light-cone quantities. While this is sometimes convenient, it can easily lead to errors. If you talk with a friend over the phone, and she says "... p-plus times ...," you will have to ask, "plus up, or plus down?" In the rest of this book we will not lower the $+$ or $-$ indices. They will always be up, and the energy will always be p^-.

We can check that the identification of p^- as light-cone energy fits together nicely with the intuition that we have developed for light-cone velocity. To this end, we confirm that a particle with small light-cone velocity also has small light-cone energy. Suppose we have

a particle moving very fast in the $+x^1$ direction. As discussed below (2.66), its light-cone velocity is very small. Since p^1 is very large, equation (2.67) gives

$$p^0 = \sqrt{(p^1)^2 + m^2c^2} = p^1\sqrt{1 + \frac{m^2c^2}{(p^1)^2}} \simeq p^1 + \frac{m^2c^2}{2p^1}. \tag{2.95}$$

The light-cone energy of the particle is therefore

$$p^- = \frac{1}{\sqrt{2}}(p^0 - p^1) \simeq \frac{m^2c^2}{2\sqrt{2}\,p^1}. \tag{2.96}$$

As anticipated, both the light-cone velocity and the light-cone energy decrease as p^1 increases.

2.6 Lorentz invariance with extra dimensions

If string theory is correct, we must entertain the possibility that spacetime has more than four dimensions. The number of time dimensions must be kept equal to one – it seems very difficult, if not altogether impossible, to construct a consistent theory with more than one time dimension. The extra dimensions must therefore be spatial. Can we have Lorentz invariance in worlds with more than three spatial dimensions? Yes. Lorentz invariance is a concept that admits a very natural generalization to spacetimes with additional dimensions.

We first extend the definition of the invariant interval ds^2 to incorporate the additional space dimensions. In a world with five spatial dimensions, for example, we would write

$$-ds^2 = -c^2dt^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2 + (dx^4)^2 + (dx^5)^2. \tag{2.97}$$

Lorentz transformations are then defined as the linear changes of coordinates that leave ds^2 invariant. This ensures that every inertial observer in the six-dimensional spacetime will agree on the value of the speed of light. With more dimensions, come more Lorentz transformations. While in four-dimensional spacetime we have boosts in the x^1, x^2, and x^3 directions, in this new world we have boosts along each of the five spatial dimensions. With three spatial coordinates, there are three basic spatial rotations: rotations that mix x^1 and x^2, those that mix x^1 and x^3, and finally those that mix x^2 and x^3. The equality of the number of boosts and the number of rotations is a special feature of four-dimensional spacetime. With five spatial coordinates, we have ten rotations, which is twice the number of boosts.

The higher-dimensional Lorentz invariance includes the lower-dimensional one: if nothing happens along the extra dimensions, then the restrictions of lower-dimensional Lorentz invariance apply. This is clear from (2.97). For motion that does not involve the extra dimensions, $dx^4 = dx^5 = 0$, and the expression for ds^2 reduces to that used in four dimensions.

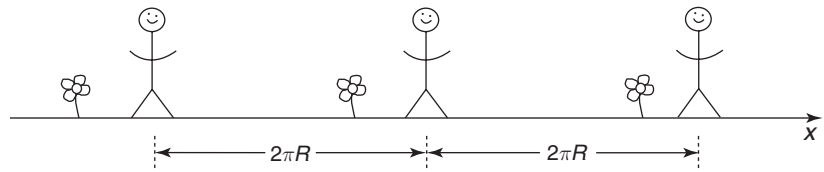

Fig. 2.4 A one-dimensional world that repeats each $2\pi R$. Several copies of Phil are shown.

2.7 Compact extra dimensions

It is possible for additional spatial dimensions to be undetected by low energy experiments if the dimensions are curled up into a compact space of small volume. In this section, we will try to understand what a compact dimension is. We will focus mainly on the case of one dimension. In Section 2.10 we will explain why small compact dimensions are hard to detect.

Consider a one-dimensional world, an infinite line, say, and let x be a coordinate along this line. For each point P along the line, there is a unique real number $x(P)$ called the x-coordinate of the point P. A good coordinate on this infinite line satisfies two conditions.

- Any two distinct points $P_1 \neq P_2$ have different coordinates: $x(P_1) \neq x(P_2)$.
- The assignment of coordinates to points is continuous: nearby points have nearly equal coordinates.

If a choice of origin is made for this infinite line, then we can use distance from the origin to define a good coordinate. The coordinate assigned to each point is the distance from that point to the origin, with a sign depending upon which side of the origin the point lies.

Imagine that you live in a world with one spatial dimension. Suppose you are walking along and notice a strange pattern: the scenery repeats each time you move a distance $2\pi R$, for some value of R. If you meet your friend Phil, you see that there are Phil clones at distances $2\pi R, 4\pi R, 6\pi R, \ldots$ down the line (see Figure 2.4). In fact, there are clones up the line, as well, with the same spacing.

There is no way to distinguish an infinite line with such a strange property from a circle with circumference $2\pi R$. Indeed, saying that this strange line is a circle *explains* the peculiar property – there really are no Phil clones; you meet the same Phil again and again as you go around the circle!

How do we express this mathematically? We can think of the circle as the open line with an *identification*. That is, we declare that points with coordinates that differ by $2\pi R$ are the same point. More precisely, two points are declared to be the same point if their coordinates differ by an integer number of $2\pi R$:

$$P_1 \sim P_2 \iff x(P_1) = x(P_2) + 2\pi R\, n, \quad n \in \mathbb{Z}. \tag{2.98}$$

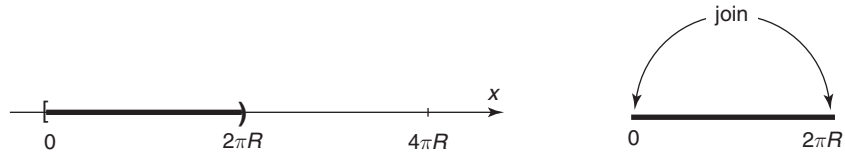

Fig. 2.5 **The interval $0 \leq x < 2\pi R$ is a fundamental domain for the line with the identification (2.99). The identified space is a circle of radius R.**

This is precise, but somewhat cumbersome, notation. With no risk of confusion, we can simply write

$$x \sim x + 2\pi R, \tag{2.99}$$

which should be read as "identify any two points whose coordinates differ by $2\pi R$." With such an identification, the open line becomes a circle. The identification has turned a non-compact dimension into a compact one. It may seem to you that a line with identifications is only a complicated way to think about a circle. We will see, however, that many physical problems become clearer when we view a compact dimension as an extended one with identifications.

The interval $0 \leq x < 2\pi R$ is a *fundamental domain* for the identification (2.99) (see Figure 2.5). A fundamental domain is a subset of the entire space that satisfies two conditions.

1. No two points in the fundamental domain are identified.
2. Any point in the entire space is in the fundamental domain or is related by the identification to some point in the fundamental domain.

Whenever possible, as we did here, the fundamental domain is chosen to be a connected region. To build the space implied by the identification, we take the fundamental domain *together* with its boundary, and implement the identifications on the boundary. In our case, the fundamental domain together with its boundary is the segment $0 \leq x \leq 2\pi R$. In this segment we identify the point $x = 0$ with the point $x = 2\pi R$. The result is the circle.

A circle of radius R can be represented in a two-dimensional plane as the set of points that are at a distance R from a point called the center of the circle. Note that the circle obtained above has been constructed directly, without the help of an embedding two-dimensional space. For our circle, there is no point, anywhere, that represents the center of the circle. We can still speak, figuratively, of the radius R of the circle, but in our case, the radius is simply the quantity which multiplied by 2π gives the total length of the circle.

On the circle, the coordinate x is no longer a good coordinate. The coordinate x is now either multivalued or discontinuous. This is a problem with any coordinate on a circle. Consider using angles to assign coordinates on the unit circle (Figure 2.6). Fix a reference point Q on the circle, and let O denote the center of the circle. To any point P on the circle we assign as a coordinate the angle $\theta(P) = \angle POQ$. This angle is naturally multivalued. The reference point Q, for example, has $\theta(Q) = 0°$ and $\theta(Q) = 360°$. If we force angles

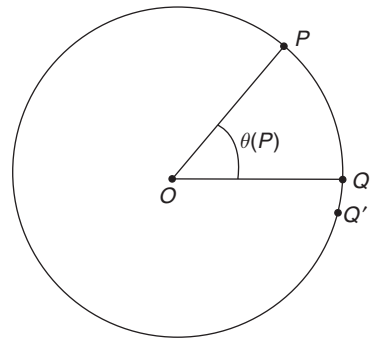

Fig. 2.6 **Using the angle θ to define a coordinate on a circle. The reference point Q is assigned zero angle: $\theta(Q) = 0$. The coordinate θ is naturally multivalued.**

to be single valued by restricting $0° \leq \theta < 360°$, for example, then they become discontinuous. Indeed, two nearby points, Q and Q', then have very different angles: $\theta(Q) = 0$, while $\theta(Q') \sim 360°$. It is easier to work with multivalued coordinates than it is to work with discontinuous ones.

If we have a world with several open dimensions, then we can apply the identification (2.99) to one of the dimensions, while doing nothing to the others. The dimension described by x turns into a circle, and the other dimensions remain open. It is possible, of course, to make more than one dimension compact. Consider, for example, the (x, y) plane, subject to *two* identifications:

$$x \sim x + 2\pi R\,, \quad y \sim y + 2\pi R. \tag{2.100}$$

It is perhaps clearer to show both coordinates simultaneously while writing the identifications. In that fashion, the two identifications are written as

$$(x, y) \sim (x + 2\pi R,\ y), \tag{2.101}$$

$$(x, y) \sim (x,\ y + 2\pi R). \tag{2.102}$$

The first identification implies that we can restrict our attention to $0 \leq x < 2\pi R$, and the second identification implies that we can restrict our attention to $0 \leq y < 2\pi R$. Thus the fundamental domain can be taken to be the square region $0 \leq x, y < 2\pi R$, as shown in Figure 2.7. To build the space implied by the identifications, we take the fundamental domain together with its boundary, forming the full square $0 \leq x, y \leq 2\pi R$, and implement the identifications on the boundary. The vertical edges are identified because they correspond to points of the form $(0, y)$ and $(2\pi R, y)$, which are identified by (2.101). The horizontal edges are identified because they correspond to points of the form $(x, 0)$ and $(x, 2\pi R)$, which are identified by (2.102). The resulting space is called a two-dimensional torus. One can visualize the torus by taking the fundamental domain (with its boundary)

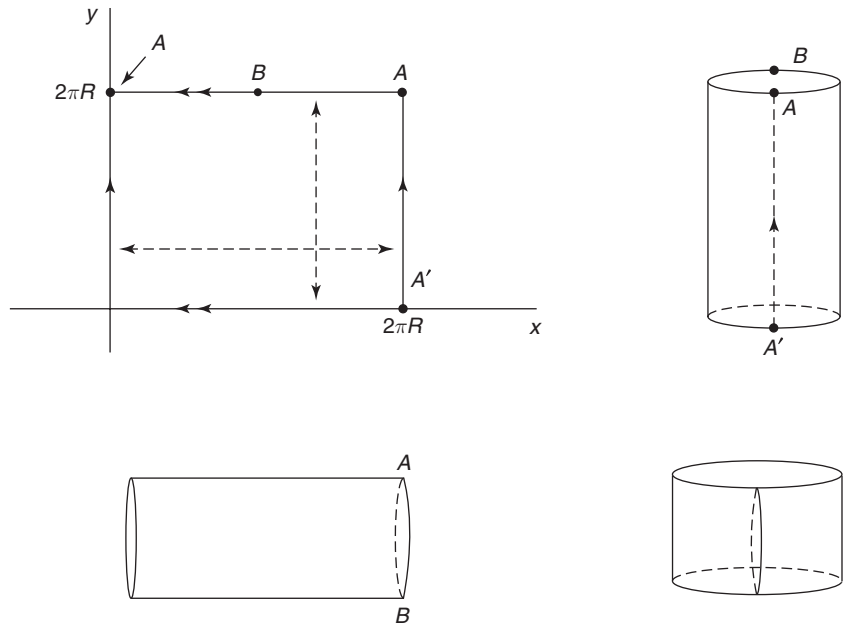

Fig. 2.7 **A square region in the plane with identifications indicated by the dashed lines and arrowheads. The resulting surface is a torus. The identification of the vertical lines gives a cylinder, shown to the right of the square region. The cylinder, shown horizontally and flattened in the bottom left, must have its edges glued to form the torus.**

and gluing the vertical edges as their identification demands. The result is a cylinder, as shown in the top right corner of Figure 2.7 (with the gluing seam dashed). In this cylinder, however, the bottom circle and the top circle must also be glued, since they are nothing other than the horizontal edges of the fundamental domain. To do this with paper, you must flatten the cylinder and then roll it up to glue the circles. The result looks like a flattened doughnut. With a flexible piece of garden hose, you could simply identify the two ends to obtain the familiar picture of a torus.

We have seen how to compactify coordinates using identifications. Some compact spaces are constructed in other ways. In string theory, however, compact spaces that arise from identifications are particularly easy to work with. We shall focus on such spaces throughout this book.

Quick calculation 2.5 Consider the plane (x, y) with the identification

$$(x\,,\ y) \sim (x + 2\pi R,\ y + 2\pi R)\,. \tag{2.103}$$

What is the resulting space? Hint: the space is most clearly exhibited using a fundamental domain for which the line $x + y = 0$ is a boundary.

2.8 Orbifolds

Sometimes identifications have fixed points, points that are related to themselves by the identification. For example, consider the real line parameterized by the coordinate x and subject to the identification $x \sim -x$. The point $x = 0$ is the unique fixed point of the identification. A fundamental domain can be chosen to be the half-line $x \geq 0$ (Figure 2.8). Note that the boundary point $x = 0$ must be included in the fundamental domain. The space obtained by the above identification is in fact the fundamental domain $x \geq 0$. This is the simplest example of an *orbifold*, a space obtained by identifications that have fixed points. An orbifold is singular at the fixed points. While the half-line $x \geq 0$ is a conventional one-dimensional manifold for $x > 0$, neighborhoods of the point $x = 0$ fail to be typical. This orbifold is called an $\mathbb{R}^1/\mathbb{Z}_2$ orbifold. Here $\mathbb{R}^1$ stands for the (one-dimensional) real line, and $\mathbb{Z}_2$ describes a basic property of the identification when it is viewed as the transformation $x \to -x$: if applied twice, it gives back the original coordinate.

Certain two-dimensional cones can be obtained as orbifolds. Begin with the (x, y) plane and identify every point with the image obtained by rotation around the origin through the angle $2\pi/N$, where $N \geq 2$ is an integer. A simple description of the identification makes use of the complex coordinate $z = x + iy$:

$$z \sim e^{\frac{2\pi i}{N}} z\,. \tag{2.104}$$

This identification does as expected: multiplication of any complex number by a phase $e^{i\alpha}$ (α real) rotates the complex number by the angle α. The identification is of $\mathbb{Z}_N$ type: viewed as the transformation $z \to e^{\frac{2\pi i}{N}} z$, if applied N times, it gives back the original coordinate. A point must be identified with all the $N - 1$ images obtained by repeated action of the transformation. The only fixed point of the $\mathbb{Z}_N$ transformation is the origin $z = 0$. A fundamental domain for (2.104), as we will explain below, is provided by the points z that satisfy the constraint

$$0 \leq \arg(z) < \frac{2\pi}{N}\,. \tag{2.105}$$

Here we recall that for $z = re^{i\theta}$, with r and θ real, we have $\arg(z) = \theta$. The fundamental domain is shown in Figure 2.9. To the right one sees the cone, obtained by gluing the rays $\arg(z) = 0$ and $\arg(z) = 2\pi/N$ using the identification (2.104). The resulting cone is called the $\mathbb{C}/\mathbb{Z}_N$ orbifold, where $\mathbb{C}$ denotes the complex plane, namely, the original two-dimensional plane equipped with a complex coordinate. The cone is singular at the apex $z = 0$ in the sense that it has concentrated curvature.

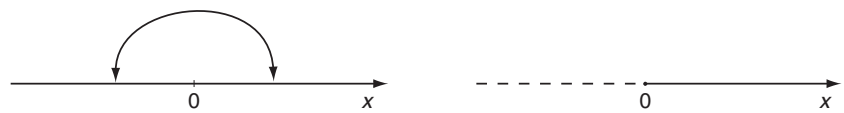

Fig. 2.8 The identification $x \sim -x$ on the real line yields the half-line. This is the $\mathbb{R}^1/\mathbb{Z}_2$ orbifold.

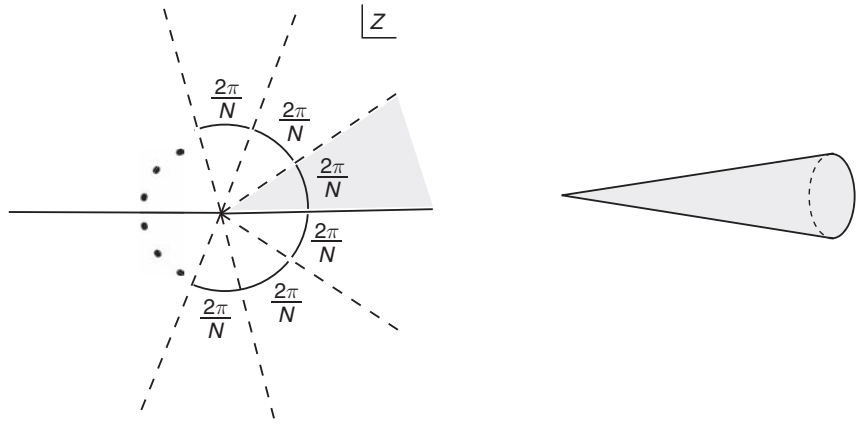

Fig. 2.9 **The fundamental domain for the identification $z \sim e^{\frac{2\pi i}{N}} z$ is shown shaded. After identification we obtain a cone – the $\mathbb{C}/\mathbb{Z}_N$ orbifold.**

Let us explain why the region defined by (2.105) is a fundamental domain. Acting repeatedly on the region with the transformation $z \to e^{\frac{2\pi i}{N}} z$, we find $N-1$ images that together with the region cover seamlessly the full complex plane. Since each point in the region has exactly $N-1$ copies, all outside the region, no two points in the region are identified. Any point in $\mathbb{C}$ that is not in the region must lie in one of its $N-1$ copies and therefore it has an image in the region. Note that our argument used the fact that N is an integer; for N irrational, for example, any point would have an infinite number of images. Our construction gives only the cones whose total angle at the apex is 2π divided by an integer. Cones with other angles exist, but they are not obtained as orbifolds. The case when N is replaced by a rational number is examined in Problem 2.7. Additional examples of orbifolds are considered in Problems 2.5, 2.6, and 2.10.

Physics on spaces with generic singularities is typically complicated and sometimes even inconsistent. Orbifolds are spaces with tractable singularities, at least as far as strings are concerned. The physics of quantum strings on an orbifold, as we will see in Chapter 13, is completely regular because the orbifold arises from identifications applied to a non-singular space where the quantum string is simple. The strings on the orbifold inherit that simplicity.

2.9 Quantum mechanics and the square well

Planck's constant $\hbar$ appears as the constant of proportionality relating the energy E and the angular frequency ω of a photon:

$$E = \hbar\omega. \tag{2.106}$$

Since ω has units of inverse time, $\hbar$ has units of energy times time. Energy has units of ML^2T^{-2} and, therefore,

$$[\hbar] = [\text{Energy}] \times [\text{Time}] = ML^2T^{-1}. \tag{2.107}$$

The value of Planck's constant is $\hbar \simeq 1.055 \times 10^{-27}$ erg $\cdot$ s.

The constant $\hbar$ appears in the basic commutation relations of quantum mechanics. The Schrödinger position and momentum operators satisfy

$$[x, p] = i\hbar. \tag{2.108}$$

If we have several spatial dimensions, then the commutation relations are

$$\left[x^i, p_j\right] = i\hbar\,\delta^i_j, \tag{2.109}$$

where the Kronecker delta is defined as in (2.32):

$$\delta^i_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases} \tag{2.110}$$

In three spatial dimensions, the indices i and j run from 1 to 3. The generalization of quantum mechanics to higher dimensions is straightforward. With d spatial dimensions, the indices in (2.109) simply run over the d possible values.

To set the stage for the analysis of a small extra dimension, we review a standard quantum mechanics problem. Consider the time-independent Schrödinger equation

$$-\frac{\hbar^2}{2m}\nabla^2\psi(x) + V(x)\psi(x) = E\,\psi(x) \tag{2.111}$$

applied to the case of a one-dimensional square-well potential of infinite height:

$$V(x) = \begin{cases} 0 & \text{if } x \in (0, a) \\ \infty & \text{if } x \notin (0, a)\,. \end{cases} \tag{2.112}$$

For $x \notin (0, a)$, the infinite potential implies $\psi(x) = 0$. In particular, $\psi(0) = \psi(a) = 0$. This is just the quantum mechanics of a particle living on a *segment*, as shown in Figure 2.10.

When $x \in (0, a)$, the Schrödinger equation becomes

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} = E\psi. \tag{2.113}$$

The solutions of (2.113), consistent with the boundary conditions, are

$$\psi_k(x) = \sqrt{\frac{2}{a}}\sin\left(\frac{k\pi x}{a}\right), \quad k = 1, 2, \ldots, \infty\,. \tag{2.114}$$

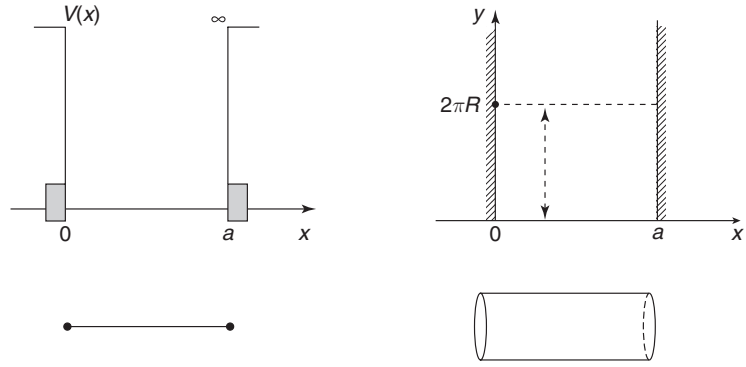

Fig. 2.10 **Left: the square-well potential in one dimension. Here the particle lives on a segment. Right: in the (x, y) plane the particle must remain within $0 < x < a$. The direction y is identified as $y \sim y + 2\pi R$. The particle lives on a cylinder.**

The value $k = 0$ is not allowed since it would make the wavefunction vanish everywhere. By performing the differentiation indicated in (2.113), we see that the energy E_k associated with the wavefunction ψ_k is

$$E_k = \frac{\hbar^2}{2m}\left(\frac{k\pi}{a}\right)^2. \tag{2.115}$$

2.10 Square well with an extra dimension

We now add an extra dimension to the square-well problem (no pun intended!). In addition to x, we include a dimension y that is curled up into a small circle of radius R. In other words, we make the identification

$$(x, y) \sim (x,\ y + 2\pi R). \tag{2.116}$$

The original dimension x has not been changed (see Figure 2.10). Since the y direction has been turned into a circle of circumference $2\pi R$, the space where the particle moves is now a *cylinder*. The cylinder has length a and circumference $2\pi R$. The potential $V(x, y)$ will remain given by (2.112) and is y-independent.

We will see that, as long as R is small, and as long as we look only at low energies, the quantum mechanics of the particle on the segment is very similar to the quantum mechanics of the particle on the cylinder. The only length scale in the original problem is the size a of the segment, so small R means $R \ll a$.

In two dimensions, the Schrödinger equation (2.111) becomes

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2}\right) = E\psi. \tag{2.117}$$

We use separation of variables in order to solve this equation. We let $\psi(x, y) = \psi(x)\,\phi(y)$ and find that the equation takes the form

$$-\frac{\hbar^2}{2m}\frac{1}{\psi(x)}\frac{d^2\psi(x)}{dx^2} - \frac{\hbar^2}{2m}\frac{1}{\phi(y)}\frac{d^2\phi(y)}{dy^2} = E. \tag{2.118}$$

The x-dependent and y-dependent terms of this equation must separately be constant, and the solutions are of the form $\psi_{k,l}(x, y) = \psi_k(x)\phi_l(y)$, where

$$\psi_k(x) = c_k \sin\left(\frac{k\pi x}{a}\right), \tag{2.119}$$

$$\phi_l(y) = a_l \sin\left(\frac{ly}{R}\right) + b_l \cos\left(\frac{ly}{R}\right). \tag{2.120}$$

The physics along the x dimension is unchanged, since the wavefunction must still vanish at the ends of the segment. Therefore (2.119) takes the same form as (2.114) and $k = 1, 2, \ldots$. The boundary condition for $\phi_l(y)$ arises from the identification $y \sim y + 2\pi R$. Since y and $y + 2\pi R$ are coordinates that represent the same point, the wavefunction must take the same value at these two arguments:

$$\phi_l(y) = \phi_l(y + 2\pi R). \tag{2.121}$$

As opposed to $\psi_k(x)$, the function $\phi_l(y)$ need not vanish for any y. As a result, the general periodic solution, recorded in (2.120), includes both sines and cosines. The presence of cosines allows a nonvanishing *constant* solution: for $l = 0$, we get $\phi_0(y) = b_0$. This solution is the key to understanding why a small extra dimension does not change the low energy physics very much.

The energy eigenvalues of the $\psi_{k,l}$ are

$$E_{k,l} = \frac{\hbar^2}{2m}\left[\left(\frac{k\pi}{a}\right)^2 + \left(\frac{l}{R}\right)^2\right]. \tag{2.122}$$

These energies correspond to doubly degenerate states when $l \neq 0$, because in this case (2.120) contains two linearly independent solutions. The extra dimension has changed the spectrum dramatically. We will see, however, that if $R \ll a$, then the *low-lying* part of the spectrum is unchanged. The rest of the spectrum changes, but these changes are not accessible *at low energies*.

Since $l = 0$ is permitted, the energy levels $E_{k,0}$ coincide with the old energy levels E_k! The new system contains all the energy levels of the old system. But it also includes additional energy levels. What is the lowest *new* energy level? To minimize the energy, each of the terms in (2.122) must be as low as possible. The minimum occurs when $k = 1$, since $k = 0$ is not allowed, and $l = 1$, since $l = 0$ gives us the old levels. The lowest new energy level is

$$E_{1,1} = \frac{\hbar^2}{2m}\left[\left(\frac{\pi}{a}\right)^2 + \left(\frac{1}{R}\right)^2\right]. \tag{2.123}$$

When $R \ll a$, the second term is much larger than the first, and

$$E_{1,1} \sim \frac{\hbar^2}{2m}\left(\frac{1}{R}\right)^2. \tag{2.124}$$

This energy is comparable to that of the level k eigenstate of the original problem (see (2.115)) when

$$\frac{k\pi}{a} \sim \frac{1}{R} \rightarrow k \sim \frac{1}{\pi}\frac{a}{R}. \tag{2.125}$$

Since R is much smaller than a, k is a very large number. So the first new energy level appears at an energy far above that of the low-lying original states. We therefore conclude that an extra dimension can remain hidden from experiments at a particular energy level as long as the dimension is small enough. Once the probing energies become sufficiently high, the effects of an extra dimension can be observed.

Curiously, the quantum mechanics of a string introduces new features. For an extra dimension *much* smaller than the already small string length ℓ_s, new low-lying states can appear! These correspond to strings that wrap around the extra dimension. They have no analog in the quantum mechanics of a point particle, and we will study them in detail in Chapter 17. In string theory, the conclusion remains true that no new low energy states arise from a small extra dimension, but there is a small qualification: the dimension must not be significantly smaller than ℓ_s. We will learn in Chapter 17 that, in string theory, the effects of a compact dimension of radius smaller than ℓ_s cannot be distinguished from those of another compact dimension with a radius larger than ℓ_s.

Problems

Problem 2.1 Exercises with units.

(a) Find the relation between coulombs (C) and esus.
(b) Explain the meaning of the unit K (degree kelvin) used for measuring temperatures, and explain its relation to the basic length, mass, and time units.
(c) Construct and evaluate a dimensionless number using the charge e of the electron (as defined in the Gaussian system of units), $\hbar$, and c. (In Heaviside–Lorentz units, the Gaussian e^2 is replaced by $\frac{e^2}{4\pi}$.)

Problem 2.2 Lorentz transformations for light-cone coordinates.

Consider coordinates $x^\mu = (x^0, x^1, x^2, x^3)$ and the associated light-cone coordinates (x^+, x^-, x^2, x^3). Write the following Lorentz transformations in terms of the light-cone coordinates.

(a) A boost with velocity parameter β in the x^1 direction.
(b) A rotation with angle θ in the x^1, x^2 plane.
(c) A boost with velocity parameter β in the x^3 direction.

Problem 2.3 Lorentz transformations, derivatives, and quantum operators.

(a) Give the Lorentz transformations for the components a_μ of a vector under a boost along the x^1 axis.

(b) Show that the objects $\frac{\partial}{\partial x^\mu}$ transform under a boost along the x^1 axis in the same way as the a_μ considered in (a) do. This checks, in a particular case, that partial derivatives with respect to upper-index coordinates x^μ behave as a four-vector with lower indices, which is why they are written as ∂_μ.

(c) Show that, in quantum mechanics, the expressions for the energy and momentum in terms of derivatives can be written compactly as $p_\mu = \frac{\hbar}{i}\frac{\partial}{\partial x^\mu}$.

Problem 2.4 Lorentz transformations as matrices.

A matrix L that satisfies (2.46) is a Lorentz transformation. Show the following.

(a) If L_1 and L_2 are Lorentz transformations so is the product $L_1 L_2$.

(b) If L is a Lorentz transformation so is the inverse matrix L^{-1}.

(c) If L is a Lorentz transformation so is the transpose matrix L^T.

Problem 2.5 Constructing simple orbifolds.

(a) Consider a circle S^1, presented as the real line with the identification $x \sim x + 2$. Choose $-1 < x \leq 1$ as the fundamental domain. The circle is the space $-1 \leq x \leq 1$ with the points $x = \pm 1$ identified. The orbifold $S^1/\mathbb{Z}_2$ is defined by imposing the (so-called) $\mathbb{Z}_2$ identification $x \sim -x$. Describe the action of this identification on the circle. Show that there are two points on the circle that are left fixed by the $\mathbb{Z}_2$ action. Find a fundamental domain for the two identifications. Describe the orbifold $S^1/\mathbb{Z}_2$ in simple terms.

(b) Consider a torus T^2, presented as the (x, y) plane with the identifications $x \sim x + 2$ and $y \sim y + 2$. Choose $-1 < x, y \leq 1$ as the fundamental domain. The orbifold $T^2/\mathbb{Z}_2$ is defined by imposing the $\mathbb{Z}_2$ identification $(x, y) \sim (-x, -y)$. Prove that there are four points on the torus that are left fixed by the $\mathbb{Z}_2$ transformation. Show that the orbifold $T^2/\mathbb{Z}_2$ is topologically a two-dimensional sphere, naturally presented as a square pillowcase with seamed edges.

Problem 2.6 Constructing the $T^2/\mathbb{Z}_3$ orbifold.

Consider the complex plane $z = x + iy$ subject to the following two identifications

$$z \sim T_1(z) = z + 1\,, \quad \text{and} \quad z \sim T_2(z) = z + e^{i\pi/3}\,.$$

(a) A fundamental domain, with its boundary, is the parallelogram with corners at $z = 0$, 1, and $e^{i\pi/3}$. Where is the fourth corner? Make a sketch and indicate the identifications on the boundary. The resulting space is an oblique torus.

(b) Consider now the *additional* $\mathbb{Z}_3$ identification

$$z \sim R(z) = e^{2\pi i/3} z\,.$$

To understand how this identification acts on the oblique torus, draw the short diagonal that divides the torus into two equilateral triangles. Describe carefully the $\mathbb{Z}_3$ action on each of the two triangles (recall that the action of R can be followed by arbitrary action with T_1, T_2, and their inverses).

(c) Determine the three fixed points of the $\mathbb{Z}_3$ action on the torus. Show that the orbifold $T^2/\mathbb{Z}_3$ is topologically a two-dimensional sphere, naturally presented as a triangular pillowcase with seamed edges and corners at the fixed points.

Problem 2.7 A more general construction for cones?

Consider the (x, y) plane and the complex coordinate $z = x + iy$. We have seen that the identification $z \sim e^{\frac{2\pi i}{N}} z$, with N an integer greater than two, can be used to construct a cone.

Examine now the identification

$$z \sim e^{2\pi i \frac{M}{N}} z\,, \quad N > M \geq 2\,,$$

where M and N are relatively prime integers (their greatest common divisor is one). One may guess that a fundamental domain is provided by the points z that satisfy $0 \leq \arg(z) < 2\pi \frac{M}{N}$. Play with low values of M and N to convince yourself that this is *not* true. Determine a fundamental domain for the identification. [Hint: use the following result. Given two relatively prime integers a and b, there exist integers m and n such that $ma + nb = 1$. Finding m and n is not easy unless you use Euclid's algorithm. Try to find, for example, integers m and n that satisfy $187m + 35n = 1$.]

Problem 2.8 Spacetime diagrams and Lorentz transformations.

Consider a spacetime diagram in which the x^0 and x^1 axes of a Lorentz frame S are represented as vertical and horizontal axes, respectively. Show that the x'^0 and x'^1 axes of the Lorentz frame S', related to S via (2.36), appear in the original spacetime diagram as oblique axes. Find the angle between the primed and unprimed axes. Show in detail how the axes appear when $\beta > 0$ and when $\beta < 0$, indicating in both cases the directions of increasing values of the coordinates.

Problem 2.9 Lightlike compactification.

The identification $x \sim x + 2\pi R$, is the statement that the coordinate x has been compactified into a circle of radius R. In this identification, the time dimension is left untouched. Consider now the strange "lightlike" compactification, in which we identify events with position and time coordinates related by

$$\begin{pmatrix} x \\ ct \end{pmatrix} \sim \begin{pmatrix} x \\ ct \end{pmatrix} + 2\pi \begin{pmatrix} R \\ -R \end{pmatrix}. \tag{1}$$

(a) Rewrite this identification using light-cone coordinates.
(b) Consider coordinates (ct', x') related to (ct, x) by a boost with velocity parameter β. Express the identifications in terms of the primed coordinates.

To interpret (1) physically, consider the family of identifications

$$\begin{pmatrix} x \\ ct \end{pmatrix} \sim \begin{pmatrix} x \\ ct \end{pmatrix} + 2\pi \begin{pmatrix} \sqrt{R^2 + R_s^2} \\ -R \end{pmatrix}, \tag{2}$$

where R_s is a length that will eventually be taken to zero, in which case (2) reduces to (1).

(c) Show that there is a boosted frame S' in which the identification (2) becomes a standard identification (i.e., the space coordinate is identified but the time coordinate is not). Find the velocity parameter of S' with respect to S and the compactification radius in this Lorentz frame S'.

(d) Represent your answer to part (c) in a spacetime diagram. Show two points related by the identification (2) and the space and time axes for the Lorentz frame S' in which the compactification is standard.

(e) Fill in the blanks in the following statement: Lightlike compactification with radius R arises by boosting a standard compactification with radius ... with Lorentz factor $\gamma \sim R/\ldots$, in the limit as $\ldots \to 0$.

Problem 2.10 A spacetime orbifold in two dimensions.

Consider a two-dimensional world with coordinates x^0 and x^1. A boost with velocity parameter β along the x^1 axis is described by the first two equations in (2.36). We want to understand the two-dimensional space that emerges if we identify

$$(x^0, x^1) \sim (x'^0, x'^1). \tag{1}$$

We are identifying spacetime points whose coordinates are related by a boost!

(a) Use the result of Problem 2.2, part (a), to recast (1) as

$$(x^+, x^-) \sim \left(e^{-\lambda} x^+ , e^{\lambda} x^-\right), \quad \text{where} \quad e^{\lambda} \equiv \sqrt{\frac{1+\beta}{1-\beta}}. \tag{2}$$

What is the range of λ? What is the orbifold fixed point? Assume now that $\beta > 0$, and thus $\lambda > 0$.

(b) Draw a spacetime diagram, indicate the x^+ and x^- axes, and sketch the family of curves

$$x^+ x^- = a^2 , \tag{3}$$

where $a > 0$ is a real constant that labels the various curves. Indicate which curves have small a and which have large a. For each value of a, equation (3) describes two disconnected curves. Show that the identification (2) relates points on each separate curve.

(c) Use the expression $-ds^2 = -2dx^+ dx^-$ for the interval to show that any curve in (3) is spacelike.

(d) Consider the two curves $x^+ x^- = a^2$ for some fixed a. The identification (2) makes each one of these curves into a circle. Find the invariant circumference of this circle by integrating the appropriate root of ds^2 between two neighboring identified points. Give your answer in terms of a and λ. Answer: $\sqrt{2}\, a\lambda$.

Roughly, as time goes from minus infinity to plus infinity, the parameter a goes from infinity down to zero and then back to infinity. This orbifold represents a universe where space is a circle. The circle begins large, contracts to zero size, and then expands again. This orbifold has one pathology: the curves $x^+ x^- = -a^2$ are actually closed timelike circles.

Problem 2.11 Extra dimension and statistical mechanics.

Write a double sum that represents the statistical mechanics partition function $Z(a, R)$ for the quantum mechanical system considered in Section 2.10. Note that $Z(a, R)$ factors as $Z(a, R) = Z(a)\tilde{Z}(R)$.

(a) Explicitly calculate $Z(a, R)$ in the very high temperature limit ($\beta = \frac{1}{kT} \to 0$). Prove that this partition function coincides with the partition function of a particle in a two-dimensional *box* with sides a and $2\pi R$. This shows that, at high temperatures, the effects of the extra dimension are visible.

(b) Assume that $R \ll a$ in such a way that there are temperatures that are large as far as the box dimension a is concerned, but small as far as the compact dimension is concerned. Write an inequality involving kT and other constants to express this possibility. Evaluate $Z(a, R)$ in this regime, but include the leading correction due to the small extra dimension.

3 Electromagnetism and gravitation in various dimensions

As a candidate theory of all interactions, string theory includes Maxwell electrodynamics and its nonlinear cousins, as well as gravitation. We review the relativistic formulation of four-dimensional electrodynamics and show how it facilitates the definition of electrodynamics in other dimensions. We give a brief description of Einstein's gravity and use the Newtonian limit to discuss the relation between Planck's length and the gravitational constant in various dimensions. We study the effect of compactification on the gravitational constant and explain how large extra dimensions could escape detection.

3.1 Classical electrodynamics

Unlike Newtonian mechanics, classical electrodynamics is a relativistic theory. In fact, Einstein was led by considerations of electrodynamics to formulate the special theory of relativity. Electrodynamics has a particularly elegant formulation in which the relativistic character of the theory is manifest. This relativistic formulation allows a natural extension of the theory to higher dimensions. Before we discuss the relativistic formulation we must review Maxwell's equations. These equations describe the dynamics of electric and magnetic fields.

Although most undergraduate and graduate courses in electrodynamics nowadays use the international system of units (SI units), the Heaviside–Lorentz system of units is far more convenient for discussions that involve relativity and extra dimensions. In this system of units, Maxwell's equations take the following form:

$$\nabla \times \vec{E} = -\frac{1}{c}\frac{\partial \vec{B}}{\partial t}, \tag{3.1}$$

$$\nabla \cdot \vec{B} = 0, \tag{3.2}$$

$$\nabla \cdot \vec{E} = \rho, \tag{3.3}$$

$$\nabla \times \vec{B} = \frac{1}{c}\vec{j} + \frac{1}{c}\frac{\partial \vec{E}}{\partial t}. \tag{3.4}$$

The above equations imply that $\vec{E}$ and $\vec{B}$ are measured with the *same* units. The first two equations are the source-free Maxwell equations. The second two involve sources: the charge density ρ, with units of charge per unit volume, and the current density $\vec{j}$, with

units of current per unit area. The Lorentz force law, which gives the rate of change of the relativistic momentum of a charged particle in an electromagnetic field, takes the form

$$\frac{d\vec{p}}{dt} = q\left(\vec{E} + \frac{\vec{v}}{c} \times \vec{B}\right). \tag{3.5}$$

Since the magnetic field $\vec{B}$ is divergenceless, it can be written as the curl of a vector, the well known vector potential $\vec{A}$:

$$\vec{B} = \nabla \times \vec{A}. \tag{3.6}$$

In electrostatics the electric field $\vec{E}$ has zero curl, and it is therefore written as (minus) the gradient of a scalar, the well known scalar potential Φ. In electrodynamics, as equation (3.1) indicates, the curl of $\vec{E}$ is not always zero. Substituting (3.6) into (3.1), we find a linear combination of $\vec{E}$ and of the time derivative of $\vec{A}$ that has zero curl:

$$\nabla \times \left(\vec{E} + \frac{1}{c}\frac{\partial \vec{A}}{\partial t}\right) = 0. \tag{3.7}$$

The object inside parentheses is set equal to $-\nabla\Phi$, and the electric field $\vec{E}$ can be written in terms of the scalar potential and the vector potential:

$$\vec{E} = -\frac{1}{c}\frac{\partial \vec{A}}{\partial t} - \nabla\Phi. \tag{3.8}$$

While the potentials $(\Phi, \vec{A})$ introduced above seem to be just auxiliary quantities used to represent electric and magnetic fields, we learn in quantum mechanics that, in fact, the potentials are more fundamental than the $\vec{E}$ and $\vec{B}$ fields. The Hamiltonian that describes the motion of a charged particle uses the potentials, not the fields. It is therefore relevant to examine possible ambiguities in the definition of the potentials. As we now show, the potentials associated with a set of $\vec{E}$ and $\vec{B}$ fields are not unique.

If we change $\vec{A}$ into $\vec{A}' = \vec{A} + \nabla\epsilon$, where ϵ is an arbitrary function of space and time, the new magnetic field $\vec{B}'$ is equal to the old one:

$$\vec{B}' = \nabla \times \vec{A}' = \nabla \times A + \nabla \times \nabla\epsilon = \vec{B}, \tag{3.9}$$

noting that the curl of a gradient is zero. The change in $\vec{A}$ would not leave $\vec{E}$ unchanged, as it is clear from (3.8). We can repair this, however, by having Φ change too. In fact, letting

$$\begin{aligned} \Phi &\longrightarrow \Phi' = \Phi - \frac{1}{c}\frac{\partial \epsilon}{\partial t}, \\ \vec{A} &\longrightarrow \vec{A}' = \vec{A} + \nabla\epsilon, \end{aligned} \tag{3.10}$$

neither $\vec{B}$ nor $\vec{E}$ is changed. The changes of the potentials indicated above are called *gauge transformations* and ϵ is the gauge parameter.

Quick calculation 3.1 Verify that $\vec{E}$, as given in (3.8), is invariant under the gauge transformations (3.10).

Two sets of potentials $(\Phi, \vec{A})$ and $(\Phi', \vec{A}')$ that are related by gauge transformations are *physically* equivalent. It follows that physically equivalent sets of the potentials give

identical electric and magnetic fields. It can happen, however, that potentials $(\Phi, \vec{A})$ and $(\Phi', \vec{A}')$ give the same electric and magnetic fields, but still one cannot find an ϵ such that (3.10) holds. In such case, the potentials are not gauge equivalent and must be considered physically different, even if their $\vec{E}$ and $\vec{B}$ fields are the same! This surprising situation can occur in spacetimes with compact spatial dimensions and will feature in our later studies of D-branes (Section 18.3). It does not happen in Minkowski space.

In the presence of compact spatial dimensions a related subtlety occurs. Given some $\vec{E}$ and $\vec{B}$ there may not exist potentials Φ and $\vec{A}$ that satisfy (3.6) and (3.8) and are well defined throughout the compact part of the space. The gauge transformations come to our help. It is not strictly necessary to have uniquely defined potentials $(\Phi, \vec{A})$ all over the compact space. A set of potentials defined on patches that cover fully the compact space is *admissible* if in the regions of overlap between any two patches the corresponding potentials are related by gauge transformations. Given our statement that potentials are needed in quantum mechanics, we must conclude that a configuration of $\vec{E}$ and $\vec{B}$ fields that does not arise from admissible potentials cannot be discussed.

By the introduction of potentials, the source-free Maxwell equations (3.1) and (3.2) are automatically satisfied. Equations (3.3) and (3.4) contain additional information. They are used to derive equations for Φ and $\vec{A}$.

3.2 Electromagnetism in three dimensions

What is electromagnetism in three spacetime dimensions? One way to produce a theory of electromagnetism in three dimensions is to begin with the four-dimensional theory and eliminate one spatial coordinate. This procedure is called dimensional reduction.

In four spacetime dimensions, both electric and magnetic fields have three spatial components: (E_x, E_y, E_z) and (B_x, B_y, B_z), respectively. It may seem likely that a reduction to a world without a z coordinate would require dropping the z components from the two fields. Surprisingly, this does not work! Maxwell's equations and the Lorentz force law make it impossible.

In order to construct a consistent three-dimensional theory, we must ensure that the dynamics does not depend on the z direction, the direction that we want to eliminate. If there is motion, it must remain restricted to the (x, y) plane. It is thus natural to require that *no quantity should have z-dependence*. This does *not* necessarily mean dropping quantities with a z index.

The Lorentz force law (3.5) is a useful guide to the construction of the lower-dimensional theory. Suppose that there is no magnetic field. Then, in order to keep the z component of momentum equal to zero we must have $E_z = 0$; the z component of the electric field must go. The case of the magnetic field is more surprising. Assume that the electric field is zero. If the velocity of the particle is a vector in the (x, y) plane, a component of the magnetic field in the plane would generate, via the cross product, a force in the z direction. On the other hand, a z component of the magnetic field would generate a force in the (x, y) plane!

We conclude that B_x and B_y must be set equal to zero, while we can keep B_z. All in all,

$$E_z = B_x = B_y = 0. \tag{3.11}$$

The left-over fields E_x, E_y, and B_z can only depend on x and y. In the three-dimensional world with coordinates t, x, and y, the z index of B_z is not a vector index. Therefore, in this reduced world, B_z behaves like a Lorentz scalar (more precisely, it is an object called a pseudo-scalar). In summary, we have a two-dimensional vector $\vec{E}$ and a scalar field B_z.

We can test the consistency of this truncation by taking a look at the x and y components of (3.1):

$$\begin{aligned} \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} &= -\frac{1}{c}\frac{\partial B_x}{\partial t}, \\ \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} &= -\frac{1}{c}\frac{\partial B_y}{\partial t}. \end{aligned} \tag{3.12}$$

Since the right-hand sides are set to zero by our truncation, the left-hand sides should vanish as well. Indeed, they do. Each term on the left-hand sides equals zero, either because it contains an E_z, or because it contains a z derivative. You may examine the consistency of the remaining equations in Problem 3.3.

While setting up three-dimensional electrodynamics was not too difficult, it is much harder to guess what five-dimensional electrodynamics should be. As we will see next, the manifestly relativistic formulation of Maxwell's equations immediately gives the appropriate generalization to other dimensions.

3.3 Manifestly relativistic electrodynamics

In the relativistic formulation of Maxwell's equations neither the electric field nor the magnetic field becomes part of a four-vector. Rather, a four-vector is obtained by combining the scalar potential Φ with the vector potential $\vec{A}$:

$$A^\mu = \left(\Phi, A^1, A^2, A^3\right). \tag{3.13}$$

The corresponding object with down indices is

$$A_\mu = \left(-\Phi, A^1, A^2, A^3\right). \tag{3.14}$$

From A_μ, we create an object known as the electromagnetic *field strength* $F_{\mu\nu}$:

$$F_{\mu\nu} \equiv \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{3.15}$$

Here $\partial_\mu \equiv \frac{\partial}{\partial x^\mu}$. Equation (3.15) implies that $F_{\mu\nu}$ is antisymmetric:

$$F_{\mu\nu} = -F_{\nu\mu}. \tag{3.16}$$

It follows from this property that all diagonal components of $F_{\mu\nu}$ vanish:

$$F_{00} = F_{11} = F_{22} = F_{33} = 0. \tag{3.17}$$

Let us calculate a few entries in $F_{\mu\nu}$. Let i denote a spatial index, that is, an index that can take the values 1, 2, and 3. Making use of (3.15) and (3.8), we find

$$F_{0i} = \frac{\partial A_i}{\partial x^0} - \frac{\partial A_0}{\partial x^i} = \frac{1}{c}\frac{\partial A^i}{\partial t} + \frac{\partial \Phi}{\partial x^i} = -E_i. \tag{3.18}$$

Similarly, we can calculate F_{12}:

$$F_{12} = \partial_1 A_2 - \partial_2 A_1 = \partial_x A_y - \partial_y A_x = B_z, \tag{3.19}$$

since $\vec{B} = \nabla \times \vec{A}$. Continuing in this manner, we can compute all the entries in the matrix $F_{\mu\nu}$:

$$F_{\mu\nu} = \begin{pmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & B_z & -B_y \\ E_y & -B_z & 0 & B_x \\ E_z & B_y & -B_x & 0 \end{pmatrix}. \tag{3.20}$$

We see that the electric and magnetic fields $\vec{E}$ and $\vec{B}$ are encoded in the field strength $F_{\mu\nu}$.

The gauge transformations (3.10) discussed before can be nicely summarized with index notation as

$$A_\mu \longrightarrow A'_\mu = A_\mu + \partial_\mu \epsilon. \tag{3.21}$$

Here A_μ and A'_μ are the gauge related potentials and, as before, the gauge parameter $\epsilon(x)$ is an arbitrary function of the spacetime coordinates.

Quick calculation 3.2 Verify that the gauge transformations (3.10) are correctly summarized by (3.21).

Since gauge transformations leave the $\vec{E}$ and $\vec{B}$ fields invariant, the field strength $F_{\mu\nu}$ must be gauge *invariant*. Indeed, we readily verify that

$$\begin{aligned} F_{\mu\nu} \longrightarrow F'_{\mu\nu} &\equiv \partial_\mu A'_\nu - \partial_\nu A'_\mu \\ &= \partial_\mu (A_\nu + \partial_\nu \epsilon) - \partial_\nu (A_\mu + \partial_\mu \epsilon) \\ &= F_{\mu\nu} + \partial_\mu \partial_\nu \epsilon - \partial_\nu \partial_\mu \epsilon \\ &= F_{\mu\nu}. \end{aligned} \tag{3.22}$$

In the last step we noted that partial derivatives commute.

Recall that the use of potentials to represent $\vec{E}$ and $\vec{B}$ automatically solves the source-free Maxwell equations (3.1) and (3.2). How are these equations written in terms of the field strength $F_{\mu\nu}$? They must be written so that they hold when (3.15) holds. Consider the following combination of field strengths:

$$T_{\lambda\mu\nu} \equiv \partial_\lambda F_{\mu\nu} + \partial_\mu F_{\nu\lambda} + \partial_\nu F_{\lambda\mu}. \tag{3.23}$$

$T_{\lambda\mu\nu}$ vanishes identically on account of (3.15):

$$\partial_\lambda \left(\partial_\mu A_\nu - \partial_\nu A_\mu\right) + \partial_\mu \left(\partial_\nu A_\lambda - \partial_\lambda A_\nu\right) + \partial_\nu \left(\partial_\lambda A_\mu - \partial_\mu A_\lambda\right) = 0, \tag{3.24}$$

using the commutativity of partial derivatives. The vanishing of $T_{\lambda\mu\nu}$,

$$\partial_\lambda F_{\mu\nu} + \partial_\mu F_{\nu\lambda} + \partial_\nu F_{\lambda\mu} = 0, \tag{3.25}$$

is a set of differential equations for the field strength. These equations are precisely the source-free Maxwell equations. To make this clear, first note that $T_{\lambda\mu\nu}$ satisfies the antisymmetry conditions

$$T_{\lambda\mu\nu} = -T_{\mu\lambda\nu}, \quad T_{\lambda\mu\nu} = -T_{\lambda\nu\mu}. \tag{3.26}$$

These two equations follow from (3.23) and the antisymmetry property $F_{\mu\nu} = -F_{\nu\mu}$ of the field strength. They state that T changes sign under the transposition of any two adjacent indices.

Quick calculation 3.3 Verify the equations in (3.26).

Any object, with however many indices, that changes sign under the transposition of every pair of adjacent indices, will change sign under the transposition of *any* two indices: to exchange any two indices you need an odd number of transpositions of adjacent indices (do you see why?). An object that changes sign under the transposition of any two indices is said to be *totally* antisymmetric. Therefore T is totally antisymmetric.

Since T is totally antisymmetric, it vanishes when any two of its indices take the same value. T is nonvanishing only when each of its three indices takes a different value. In such case, different orderings of these three fixed values will give T components that can differ at most by a sign. Since we are setting T to zero these various orderings do not give new conditions. Because we have four spacetime coordinates, selecting three different indices can only be done in four different ways – leaving out a different index each time. Thus the vanishing of T gives four nontrivial equations. These four equations are the three components of equation (3.1) and equation (3.2). The vanishing of T_{012}, for example, gives us

$$\partial_0 F_{12} + \partial_1 F_{20} + \partial_2 F_{01} = \frac{1}{c}\frac{\partial B_z}{\partial t} + \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} = 0. \tag{3.27}$$

This is the z component of equation (3.1). The other three choices of indices lead to the remaining three equations (Problem 3.2).

How can we describe Maxwell equations (3.3) and (3.4) in our present framework? Since these equations have sources, we must introduce a *current* four-vector:

$$j^\mu = \left(c\rho, j^1, j^2, j^3\right), \tag{3.28}$$

where ρ is the charge density and $\vec{j} = (j^1, j^2, j^3)$ is the current density. In addition, we raise the indices of the field tensor to obtain the field tensor with upper indices:

$$F^{\mu\nu} = \eta^{\mu\alpha}\eta^{\nu\beta}F_{\alpha\beta}. \tag{3.29}$$

Quick calculation 3.4 Show that

$$F^{\mu\nu} = -F^{\nu\mu}, \quad F^{0i} = -F_{0i}, \quad F^{ij} = F_{ij}. \tag{3.30}$$

Equation (3.29), together with the definition of $F_{\mu\nu}$, gives

$$F^{\mu\nu} = \eta^{\mu\alpha}\eta^{\nu\beta}(\partial_\alpha A_\beta - \partial_\beta A_\alpha) = \eta^{\mu\alpha}\partial_\alpha(\eta^{\nu\beta}A_\beta) - \eta^{\nu\beta}\partial_\beta(\eta^{\mu\alpha}A_\alpha), \tag{3.31}$$

where the constancy of the metric components was used to move them across the derivatives. It is customary to apply the rules for raising and lowering indices to partial derivatives, so we write $\partial^\mu \equiv \eta^{\mu\alpha}\partial_\alpha$. As a result,

$$F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu. \tag{3.32}$$

It follows from (3.30) and (3.20) that

$$F^{\mu\nu} = \begin{pmatrix} 0 & E_x & E_y & E_z \\ -E_x & 0 & B_z & -B_y \\ -E_y & -B_z & 0 & B_x \\ -E_z & B_y & -B_x & 0 \end{pmatrix}. \tag{3.33}$$

Using this equation and the current vector (3.28), we can encapsulate Maxwell's equations (3.3) and (3.4) as (Problem 3.2)

$$\frac{\partial F^{\mu\nu}}{\partial x^\nu} = \frac{1}{c}\, j^\mu. \tag{3.34}$$

In the absence of sources this equation becomes

$$\partial_\nu F^{\mu\nu} = 0 \longrightarrow \partial_\nu\partial^\mu A^\nu - \partial^2 A^\mu = 0, \tag{3.35}$$

where we have written $\partial^2 = \partial^\mu\partial_\mu$.

Equations (3.15) together with equations (3.34) are equivalent to Maxwell's equations in four dimensions. We will take these equations to *define* Maxwell theory in arbitrary dimensions. In d spatial dimensions the Lorentz vector A^μ has components $(\Phi, \vec{A})$ where $\vec{A}$ is a d-dimensional spatial vector.

In three-dimensional spacetime, for example, the matrix $F_{\mu\nu}$ is a 3-by-3 antisymmetric matrix, obtained from (3.20) by discarding the last row and the last column:

$$F_{\mu\nu} = \begin{pmatrix} 0 & -E_x & -E_y \\ E_x & 0 & B_z \\ E_y & -B_z & 0 \end{pmatrix}. \tag{3.36}$$

This immediately reproduces the main result of Section 3.2; B_x, B_y, and E_z are to be set to zero.

Motivated by (3.33), in arbitrary dimensions we will call F^{0i} the electric field E_i:

$$E_i \equiv F^{0i} = -F_{0i}. \tag{3.37}$$

The electric field is a spatial vector. Equation (3.18) implies that, in any number of dimensions,

$$\vec{E} = -\frac{1}{c}\frac{\partial \vec{A}}{\partial t} - \nabla\Phi. \tag{3.38}$$

The magnetic field is identified with the F^{ij} components of the field strength. In four-dimensional spacetime F^{ij} is a 3-by-3 antisymmetric matrix. Its three independent entries are the components of the magnetic field vector (see (3.33)). In dimensions other than four, the magnetic field is no longer a spatial vector. In three spacetime dimensions the magnetic field is a single-component object. In five spacetime dimensions the magnetic field has as many entries as a 4-by-4 antisymmetric matrix, six entries. That many components do not fit into a spatial vector.

Our next goal is the determination of the electric field produced by a point charge in a spacetime with an arbitrary but fixed number of spatial dimensions. To this end, we must first learn how to calculate the volumes of higher-dimensional spheres. We turn to this subject now.

3.4 An aside on spheres in higher dimensions

Since we want to work in various numbers of dimensions we should be precise when speaking about spheres and their volumes. When we speak loosely we tend to confuse spheres and *balls*, at least in the precise sense in which they are defined in mathematics. When you say that the volume of a sphere of radius R is $\frac{4}{3}\pi R^3$, you should really be saying that this is the volume of the *three-ball* B^3 – the three-dimensional space enclosed by the two-dimensional *two-sphere* S^2. In three-dimensional space $\mathbb{R}^3$ with coordinates x_1, x_2, and x_3, we write the three-ball as the region defined by

$$B^3(R) \; : \quad x_1^2 + x_2^2 + x_3^2 \leq R^2. \tag{3.39}$$

This region is enclosed by the two-sphere:

$$S^2(R) \; : \quad x_1^2 + x_2^2 + x_3^2 = R^2. \tag{3.40}$$

The superscripts in B or S denote the dimensionality of the space in question. When we drop the explicit argument R, we mean that $R = 1$. Lower-dimensional examples are also familiar. B^2 is a two-dimensional disk – the region enclosed in $\mathbb{R}^2$ by the one-dimensional unit radius circle S^1. In arbitrary dimensions we define balls and spheres as subspaces of $\mathbb{R}^d$:

$$B^d(R) \; : \quad x_1^2 + x_2^2 + \cdots + x_d^2 \le R^2. \tag{3.41}$$

This is the region enclosed by the sphere $S^{d-1}(R)$:

$$S^{d-1}(R) \; : \quad x_1^2 + x_2^2 + \cdots + x_d^2 = R^2. \tag{3.42}$$

One last piece of terminology: to avoid confusion we will always speak of volumes. If a space is one-dimensional we take volume to mean length. If a space is two-dimensional we take volume to mean area. All higher-dimensional spaces have just volumes. The volumes of the one- and two-dimensional spheres are

$$\begin{aligned} \text{vol}\,(S^1(R)) &= 2\pi R, \\ \text{vol}\,(S^2(R)) &= 4\pi R^2. \end{aligned} \tag{3.43}$$

Unless you have worked with other spheres before, you probably do not know what the volume of S^3 is.

Since volume has units of length to the power of the space dimension, the volume of a sphere of radius R is related to the volume of a sphere of unit radius by

$$\text{vol}\,(S^{d-1}(R)) = R^{d-1}\ \text{vol}(S^{d-1}). \tag{3.44}$$

Since the radius dependence of the volume is easily recovered, it suffices to record the volumes of unit spheres:

$$\begin{aligned} \text{vol}\,(S^1) &= 2\pi, \\ \text{vol}\,(S^2) &= 4\pi. \end{aligned} \tag{3.45}$$

Let us now begin our calculation of the volume of the sphere S^{d-1}. For this purpose consider $\mathbb{R}^d$ with coordinates $x_1, x_2, \ldots, x_d$, and let r be the radial coordinate:

$$r^2 = x_1^2 + x_2^2 + \cdots + x_d^2. \tag{3.46}$$

We will find the desired volume by evaluating in two different ways the following integral:

$$I_d = \int_{\mathbb{R}^d} dx_1 dx_2 \ldots dx_d\, e^{-r^2}. \tag{3.47}$$

First we proceed directly. Using (3.46) in the exponential factor, the integral becomes a product of d Gaussian integrals:

$$I_d = \prod_{i=1}^{d} \int_{-\infty}^{\infty} dx_i\, e^{-x_i^2} = (\sqrt{\pi})^d = \pi^{d/2}. \tag{3.48}$$

Now we proceed indirectly. We do the integral by breaking $\mathbb{R}^d$ into thin spherical shells. Since the space of constant r is the sphere $S^{d-1}(r)$, the volume of a shell lying between r and $r + dr$ equals the volume of $S^{d-1}(r)$ times dr. Therefore,

$$\begin{aligned} I_d &= \int_0^\infty dr\ \text{vol}(S^{d-1}(r))\, e^{-r^2} = \text{vol}\,(S^{d-1}) \int_0^\infty dr\, r^{d-1}\, e^{-r^2} \\ &= \frac{1}{2}\,\text{vol}\,(S^{d-1}) \int_0^\infty dt\, e^{-t}\, t^{\frac{d}{2}-1}, \end{aligned} \tag{3.49}$$

where use was made of (3.44), and in the final step we changed the variable of integration to $t = r^2$. The last integral on the right-hand side can be expressed in terms of the gamma function, a very useful special function. For positive x the gamma function $\Gamma(x)$ is defined by

$$\Gamma(x) = \int_0^\infty dt\, e^{-t} t^{x-1}, \quad x > 0. \tag{3.50}$$

Unless $x > 0$ the integral does not converge near $t = 0$. With this definition, equation (3.49) becomes

$$I_d = \frac{1}{2} \text{vol}\,(S^{d-1})\, \Gamma\left(\tfrac{d}{2}\right). \tag{3.51}$$

Comparing with the earlier evaluation (3.48), we get our final result:

$$\text{vol}\,(S^{d-1}) = \frac{2\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)}. \tag{3.52}$$

It now remains to calculate the value of $\Gamma(d/2)$. Since d is an integer, we must determine the values of the gamma function for both integer and half-integer arguments. To find $\Gamma(1/2)$ we use the definition (3.50) and let $t = u^2$:

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty dt\, e^{-t}\, t^{-1/2} = 2\int_0^\infty du\, e^{-u^2} = \sqrt{\pi}. \tag{3.53}$$

Similarly,

$$\Gamma(1) = \int_0^\infty dt\, e^{-t} = 1. \tag{3.54}$$

For larger arguments the calculation of the gamma function is simplified using a recursion relation. To obtain this relation, begin with

$$\Gamma(x+1) = \int_0^\infty dt\, e^{-t} t^x, \quad x > 0\,, \tag{3.55}$$

which can be rewritten as

$$\Gamma(x+1) = -\int_0^\infty dt \left(\frac{d}{dt} e^{-t}\right) t^x = -\int_0^\infty dt \left(\frac{d}{dt}(e^{-t}t^x) - x e^{-t} t^{x-1}\right). \tag{3.56}$$

The boundary terms vanish for $x > 0$ and we find that

$$\Gamma(x+1) = x\,\Gamma(x), \quad x > 0. \tag{3.57}$$

Using this recursion relation we find, for example,

$$\Gamma\left(\frac{3}{2}\right) = \frac{1}{2}\cdot\Gamma\left(\frac{1}{2}\right) = \frac{1}{2}\sqrt{\pi}, \quad \Gamma\left(\frac{5}{2}\right) = \frac{3}{2}\cdot\Gamma\left(\frac{3}{2}\right) = \frac{3}{4}\sqrt{\pi}. \tag{3.58}$$

For integer arguments the gamma function is related to the factorial:

$$\Gamma(5) = 4\cdot\Gamma(4) = 4\cdot 3\cdot\Gamma(3) = 4\cdot 3\cdot 2\cdot\Gamma(2) = 4\cdot 3\cdot 2\cdot 1\cdot\Gamma(1) = 4!.$$

Therefore, for $n \in \mathbb{Z}$ and $n \geq 1$, we have

$$\Gamma(n) = (n-1)!, \tag{3.59}$$

where we recall that $0! = 1$. We can now test our formula (3.52) in the familiar cases:

$$\begin{aligned} \text{vol}\,(S^1) = \text{vol}\,(S^{2-1}) &= \frac{2\pi}{\Gamma(1)} = 2\pi, \\ \text{vol}\,(S^2) = \text{vol}\,(S^{3-1}) &= \frac{2\pi^{3/2}}{\Gamma\left(\frac{3}{2}\right)} = 4\pi, \end{aligned} \tag{3.60}$$

in agreement with the known values. For the less familiar S^3 we find

$$\text{vol}\,(S^3) = \text{vol}\,(S^{4-1}) = \frac{2\pi^2}{\Gamma(2)} = 2\pi^2. \tag{3.61}$$

Quick calculation 3.5 Show that $\text{vol}\,(B^d) = \pi^{d/2}/\,\Gamma\left(1 + \frac{d}{2}\right)$.

3.5 Electric fields in higher dimensions

In this section we calculate the electric field due to a point charge in a world with d spatial dimensions. Here d could be three, in which case the answer is familiar, or less than three, but we are particularly interested in $d > 3$. To do this calculation we will use the general version of Maxwell's equations appropriate for an arbitrary number of spatial dimensions. As you may imagine, the electric field of a point charge is radial. Our calculation will give the radial dependence and the normalization of the electric field. With minor modifications, this result will also inform us about the gravitational fields of point particles in d spatial dimensions.

Our computation is based on the zeroth component of equation (3.34):

$$\frac{\partial F^{0i}}{\partial x^i} = \rho. \tag{3.62}$$

Since $F^{0i} = E_i$ (see (3.37)), this equation is just Gauss' law:

$$\nabla \cdot \vec{E} = \rho. \tag{3.63}$$

Gauss' law is valid in all dimensions! Equation (3.63) can be used to determine the electric field of a point charge. Let us first review how this is done in the familiar setting of three spatial dimensions.

Consider a point charge q, a two-sphere $S^2(r)$ of radius r centered on the charge, and the three-ball $B^3(r)$ whose boundary is the two-sphere. We integrate both sides of equation (3.63) over the three-ball to find

$$\int_{B^3} d(\text{vol})\, \nabla \cdot \vec{E} = \int_{B^3} d(\text{vol})\, \rho. \tag{3.64}$$

We use the divergence theorem on the left-hand side and note that the volume integral on the right-hand side gives the total charge:

$$\int_{S^2(r)} \vec{E} \cdot d\vec{a} = q\,. \tag{3.65}$$

Since the magnitude $E(r)$ of $\vec{E}$ is constant over the two-sphere, we get

$$\mathrm{vol}(S^2(r))\, E(r) = q. \tag{3.66}$$

The volume of the two-sphere is just its area $4\pi r^2$, so

$$E(r) = \frac{q}{4\pi r^2}. \tag{3.67}$$

This is the familiar result for the electric field of a point charge in three spatial dimensions. The electric field magnitude falls off like $1/r^2$.

For dimensions higher than three, the starting point (3.63) is good, so we must ask if the divergence theorem also holds. It turns out that it does. We will first state the theorem in d spatial dimensions, and then we will give some justification for it.

Consider a d-dimensional subspace V^d of $\mathbb{R}^d$ and let ∂V^d denote the boundary of V^d. Moreover, let $\vec{E}$ be a vector field in $\mathbb{R}^d$. The divergence theorem states that

$$\int_{V^d} d(\mathrm{vol})\, \nabla \cdot \vec{E} = \text{Flux of } \vec{E} \text{ across } \partial V^d = \int_{\partial V^d} \vec{E} \cdot d\vec{v}. \tag{3.68}$$

The last right-hand side requires some explanation. At any point on ∂V^d, the space ∂V^d is locally approximated by the $(d-1)$-dimensional tangent hyperplane. For a small piece of ∂V^d around this point, the associated vector $d\vec{v}$ is a vector orthogonal to the hyperplane, pointing out of the volume, and with magnitude equal to the volume of the small piece under consideration. Note that this explanation is in accord with your experience in $\mathbb{R}^3$, where $d\vec{v}$ corresponds to the area vector element $d\vec{a}$.

Let us justify the divergence theorem for the case of four space dimensions. Following a strategy used in elementary textbooks, it suffices to prove the divergence theorem for a small hypercube – the result for general subspaces follows by breaking such spaces into many small hypercubes. Because it is not easy to imagine a four-dimensional hypercube, we might as well use a three-dimensional picture with four-dimensional labels (Figure 3.1). We use Cartesian coordinates x, y, z, w, and consider a cube whose faces lie on hyperplanes selected by the condition that one of the coordinates is constant. Let one face of the cube and the face opposite to it lie on hyperplanes of constant x and constant $x + dx$, respectively. The outgoing normal vectors are $\vec{e}_x$, for the face at $x + dx$, and $(-\vec{e}_x)$, for the face at x. The volume of each of these two faces equals $dydzdw$, where dy, dz, and dw, together with dx, are the lengths of the edges of the cube. For an arbitrary electric field $\vec{E}(x, y, z, w)$, only the x component contributes to the flux through these two faces. The contribution is

$$[\, E_x(x + dx, y, z, w) - E_x(x, y, z, w)\,]\, dydzdw \simeq \frac{\partial E_x}{\partial x}\, dxdydzdw. \tag{3.69}$$

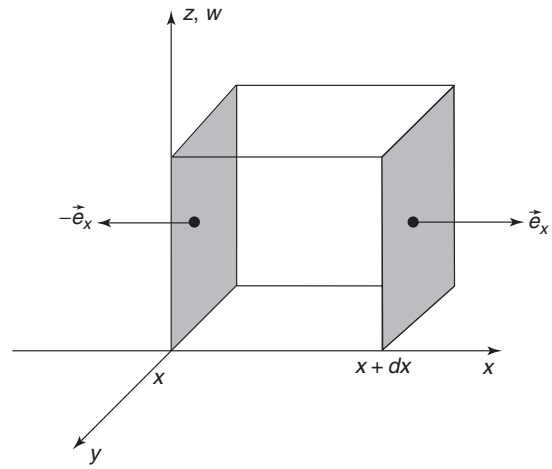

Fig. 3.1 **An attempt at a representation of a four-dimensional hypercube. The two faces of constant *x* are shown (shaded) together with their outgoing normal vectors.**

Analogous expressions hold for the flux across the three other pairs of faces. The total net flux from the little cube is just

$$\text{Flux of } \vec{E} = \left(\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} + \frac{\partial E_w}{\partial w}\right) dx\,dy\,dz\,dw = \nabla \cdot \vec{E}\, d\,(\text{vol}). \tag{3.70}$$

This result is precisely the divergence theorem (3.68) applied to an infinitesimal hypercube. This is what we wanted to show.

We can now return to the computation of the electric field due to a point charge in a world with d spatial dimensions. Consider a point charge q, the sphere $S^{d-1}(r)$ of radius r centered on the charge (this is the sphere that surrounds the charge), and the ball $B^d(r)$ whose boundary is the sphere $S^{d-1}(r)$. Again, we integrate both sides of equation (3.63) over the ball $B^d(r)$:

$$\int_{B^d} d\,(\text{vol}) \nabla \cdot \vec{E} = \int_{B^d} d\,(\text{vol})\, \rho. \tag{3.71}$$

The volume integral on the right-hand side gives the total charge, and the divergence theorem (3.68) relates the left-hand side to a flux integral:

$$\text{Flux of } \vec{E} \text{ across } S^{d-1}(r) = q. \tag{3.72}$$

The flux equals the magnitude of the electric field times the volume of $S^{d-1}(r)$, so

$$E(r)\,\text{vol}\,(S^{d-1}(r)) = q. \tag{3.73}$$

Making use of (3.52) we find

$$E(r) = \frac{\Gamma\left(\frac{d}{2}\right)}{2\pi^{d/2}} \frac{q}{r^{d-1}}. \tag{3.74}$$

This is the value of the electric field for a point charge in a world with d spatial dimensions. For $d = 3$ we recover the inverse-squared dependence of the electric field. In higher dimensions the electric field falls off faster at large distances. For each additional spatial dimension we get an additional factor of $1/r$ in the radial dependence of the electric field.

Quick calculation 3.6 Verify that for $d = 3$ equation (3.74) coincides with (3.67).

Quick calculation 3.7 The force $\vec{F}$ on a test charge q in an electric field $\vec{E}$ is $\vec{F} = q\vec{E}$. What are the units of charge in various dimensions?

The electrostatic potential Φ is also of interest. For time independent fields (3.38) gives

$$\vec{E} = -\nabla\Phi. \tag{3.75}$$

This equation, together with Gauss' law, gives the Poisson equation:

$$\nabla^2\Phi = -\rho, \tag{3.76}$$

which can be used to calculate the potential due to a charge distribution. The two equations above hold in all dimensions using, of course, the appropriate definitions of the gradient and the Laplacian.

3.6 Gravitation and Planck's length

Einstein's theory of general relativity is a theory of gravitation. In this very elegant theory the dynamical variables encode the geometry of spacetime. When gravitational fields are sufficiently weak and velocities are small, Newtonian gravitation is accurate enough, and one need not work with the more complex machinery of general relativity. We can use Newtonian gravity to understand the definition of Planck's length in various dimensions and its relation to the gravitational constant. These are interesting issues that we will explain here and in the rest of the present chapter. Nevertheless, when gravitation emerges in string theory, it does so in the language of Einstein's theory of general relativity. To be able to recognize the appearance of gravity among the quantum vibrations of the relativistic string you need a little familiarity with the language of general relativity. Here you will take a first look at the concepts involved in this remarkable theory.

Most physicists do not expect general relativity to hold at truly small distances nor for extremely large gravitational fields. This is a realm where string theory, the first serious candidate for a quantum theory of gravitation, is necessary. General relativity is the large-distance/weak-gravity limit of string theory. String theory *modifies* general relativity; it must do so to make it consistent with quantum mechanics. The conceptual framework which underlies these modifications is not clear yet. It will no doubt emerge as we understand string theory better in the years to come.

The spacetime of special relativity, Minkowski spacetime, is the arena for physics in the *absence* of gravitational fields. The geometrical properties of Minkowski spacetime are

encoded by the metric formula (2.21), which gives the invariant interval separating two nearby events:

$$-ds^2 = \eta_{\mu\nu} dx^\mu dx^\nu. \tag{3.77}$$

Here the Minkowski metric $\eta_{\mu\nu}$ is a constant metric, represented as a matrix with entries $(-1, 1, \ldots, 1)$ along the diagonal. Minkowski space is said to be a flat space. In the presence of a gravitational field, the metric becomes dynamical. We then write

$$-ds^2 = g_{\mu\nu}(x) dx^\mu dx^\nu, \tag{3.78}$$

where the constant $\eta_{\mu\nu}$ is replaced by the metric $g_{\mu\nu}(x)$. If there is a gravitational field, the metric is in general a nontrivial function of the spacetime coordinates. The metric $g_{\mu\nu}$ is defined to be symmetric

$$g_{\mu\nu}(x) = g_{\nu\mu}(x). \tag{3.79}$$

It is also customary to define $g^{\mu\nu}(x)$ as the inverse of the $g_{\mu\nu}(x)$ matrix:

$$g^{\mu\alpha}(x)\, g_{\alpha\nu}(x) = \delta^\mu_\nu. \tag{3.80}$$

For many physical phenomena gravity is very weak, and the metric $g_{\mu\nu}(x)$ can be chosen to be very close to the Minkowski metric $\eta_{\mu\nu}$. We then write,

$$g_{\mu\nu}(x) = \eta_{\mu\nu} + h_{\mu\nu}(x), \tag{3.81}$$

and we view $h_{\mu\nu}(x)$ as a small fluctuation around the Minkowski metric. This expansion is done, for example, to study gravity waves. Those waves represent small "ripples" on top of the Minkowski metric. Einstein's equations for the gravitational field are written in terms of the spacetime metric $g_{\mu\nu}(x)$. These equations imply that matter or energy sources curve the spacetime manifold. For weak gravitational fields, Einstein's equations can be expanded in powers of $h_{\mu\nu}$ using (3.81). In the absence of sources, the resulting linearized equation for $h_{\mu\nu}$ is

$$\partial^2 h^{\mu\nu} - \partial_\alpha(\partial^\mu h^{\nu\alpha} + \partial^\nu h^{\mu\alpha}) + \partial^\mu \partial^\nu h = 0. \tag{3.82}$$

Here $h^{\mu\nu} \equiv \eta^{\mu\alpha}\eta^{\nu\beta} h_{\alpha\beta}$ and $h \equiv \eta^{\mu\nu} h_{\mu\nu} = -h_{00} + h_{11} + h_{22} + h_{33}$. Equation (3.82) is the gravitational analog of equation (3.35), which describes Maxwell fields in the absence of sources. While (3.35) is exact, (3.82) is only valid for weak gravitational fields. It is the linear approximation to a nonlinear equation that includes additional terms quadratic and higher order in h.

The analogy with electromagnetism extends to the existence of gauge transformations. Einstein's gravity has gauge transformations. They arise because the use of different systems of coordinates yields equivalent descriptions of gravitational physics. In learning string theory in this book you will get to appreciate the freedom to choose coordinates on the surfaces generated by moving strings. In general relativity, an infinitesimal change of coordinates

$$x^{\mu'} = x^\mu + \epsilon^\mu(x), \tag{3.83}$$

can be viewed as an infinitesimal change of the metric $g_{\mu\nu}$ and, using (3.81), as an infinitesimal change of the fluctuating field $h^{\mu\nu}$. One can show that the change is given by

$$\delta h^{\mu\nu} = \delta_0 h^{\mu\nu} + \mathcal{O}(\epsilon, h)\,, \quad \text{with} \quad \delta_0 h^{\mu\nu} \equiv \partial^\mu \epsilon^\nu + \partial^\nu \epsilon^\mu. \tag{3.84}$$

As indicated above, the infinitesimal change $\delta h^{\mu\nu}$ is given by $\delta_0 h^{\mu\nu}$ plus corrections, written as $\epsilon \cdot \mathcal{O}(h)$, that are linear in ϵ and linear in the fluctuation h itself. The invariance of the full nonlinear equation of motion under the gauge transformation $\delta h^{\mu\nu}$ requires the invariance of the linearized equation of motion (3.82) under $\delta_0 h^{\mu\nu}$. Indeed, when we vary the hs in (3.82) using $\delta_0 h^{\mu\nu}$ we get terms linear in ϵ but without an h. These terms must cancel out completely because all other variations will contain at least one field h: this is clear for the variations of (3.82) using the terms represented by $\epsilon \cdot \mathcal{O}(h)$ and for all variations of terms quadratic and higher order in h in the complete equation of motion. We will check the invariance of (3.82) under the transformation $\delta_0 h^{\mu\nu}$ in Chapter 10. In Maxwell theory the gauge parameter has no indices, but in general relativity the gauge parameter has a vector index.

As we mentioned before, Newtonian gravitation emerges from general relativity in the approximation of weak gravitational fields and motion with small velocities. For many purposes Newtonian gravity suffices. Starting now, and for the rest of this chapter, we will use Newtonian gravity to understand the definition of Planck's length in various dimensions, and to investigate how gravitational constants behave when some spatial dimensions are curled up. The results that we will obtain hold also in the full theory of general relativity.

Newton's law of gravitation in four dimensions states that the force of attraction between two masses m_1 and m_2 separated by a distance r is given by

$$|\vec{F}^{(4)}| = \frac{G\, m_1 m_2}{r^2}, \tag{3.85}$$

where G denotes the four-dimensional Newton constant. It follows that the units of the gravitational constant G are

$$[G] = [\text{Force}]\frac{L^2}{M^2} = \frac{ML}{T^2}\frac{L^2}{M^2} = \frac{L^3}{MT^2}. \tag{3.86}$$

The numerical value for the constant G is determined experimentally:

$$G = 6.674 \times 10^{-11}\, \frac{\text{m}^3}{\text{kg} \cdot \text{s}^2}. \tag{3.87}$$

Since $[c] = L/T$ and $[\hbar] = ML^2/T$, the three fundamental constants G, c, and $\hbar$ can be written as

$$G = 6.674 \times 10^{-11}\, \frac{\text{m}^3}{\text{kg} \cdot \text{s}^2}, \quad c = 2.998 \times 10^8\, \frac{\text{m}}{\text{s}}, \quad \hbar = 1.055 \times 10^{-34}\, \frac{\text{kg} \cdot \text{m}^2}{\text{s}}. \tag{3.88}$$

In the study of gravitation it is sometimes convenient to use a "Planckian" system of units. Since we have three basic units, those of length, time, and mass, we can find new units of length, time, and mass such that the three fundamental constants, G, c, and $\hbar$ take

the numerical value of *one* in those units. These units are called the Planck length ℓ_P, the Planck time t_P, and the Planck mass m_P, respectively. In those units

$$G = 1 \cdot \frac{\ell_P^3}{m_P\, t_P^2}, \quad c = 1 \cdot \frac{\ell_P}{t_P}, \quad \hbar = 1 \cdot \frac{m_P\, \ell_P^2}{t_P}, \tag{3.89}$$

without additional numerical constants – as opposed to equation (3.88). The above equations allow us to solve for ℓ_P, t_P, and m_P in terms of G, c, and $\hbar$. One readily finds

$$\ell_P = \sqrt{\frac{G\hbar}{c^3}} = 1.616 \times 10^{-33}\ \text{cm}, \tag{3.90}$$

$$t_P = \frac{\ell_P}{c} = \sqrt{\frac{G\hbar}{c^5}} = 5.391 \times 10^{-44}\ \text{s}, \tag{3.91}$$

$$m_P = \sqrt{\frac{\hbar c}{G}} = 2.176 \times 10^{-5}\ \text{g}. \tag{3.92}$$

These numbers represent scales at which relativistic quantum gravity effects can be important. Indeed, the Planck length is an extremely small length, and the Planck time is an incredibly short time – the time it takes light to travel the Planck length! While Einstein's gravity can be used down to relatively small distances and back to relatively early times in the history of the universe, a quantum gravity theory (such as string theory) is needed to study gravity at distances of the order of the Planck length or to investigate the universe when it was Planck-time old.

There is an equivalent way to characterize the Planck length: ℓ_P is the unique length that can be constructed using *only* powers of G, c, and $\hbar$. One thus sets

$$\ell_P = (G)^\alpha\, (c)^\beta\, (\hbar)^\gamma, \tag{3.93}$$

and fixes the constants α, β, and γ so that the right-hand side has units of length.

Quick calculation 3.8 Show that this condition fixes uniquely $\alpha = \gamma = 1/2$, and $\beta = -3/2$, thus reproducing the result in (3.90).

It may appear that m_P is not a very large mass, but it is, in fact, a spectacularly large mass from the viewpoint of elementary particle physics. The mass m_P is roughly 10^{19} times larger than the mass of the proton. If the fundamental theory of nature is based on the basic constants G, c, and $\hbar$, it is then a great mystery why the masses of the elementary particles are so much smaller than the "obvious" mass m_P that can be built from the basic constants. This puzzle is usually called the *hierarchy* problem.

For an additional perspective on the Planck mass, consider the following question: what should be the mass M of the proton so that the gravitational force between two protons cancels the electric repulsion force between them? Equating the magnitudes of the electric and gravitational forces we get

$$\frac{GM^2}{r^2} = \frac{e^2}{4\pi r^2} \longrightarrow GM^2 = \frac{e^2}{4\pi}. \tag{3.94}$$

It is convenient to divide both sides of the equation by $\hbar c$ to find

$$\frac{GM^2}{\hbar c} = \frac{e^2}{4\pi\hbar c} \simeq \frac{1}{137} \longrightarrow \frac{M^2}{m_{\rm P}^2} \simeq \frac{1}{137}, \tag{3.95}$$

where use was made of (3.92). We thus find $M \simeq m_{\rm P}/12$, or about one-tenth of the Planck mass. The dimensionless ratio $e^2/(4\pi\hbar c)$ is called the fine structure constant. It was evaluated above using the Heaviside–Lorentz definition of electric charge where $e = \sqrt{4\pi}\, 4.8 \times 10^{-10}$ esu (see Problem 2.1(c)).

Quick calculation 3.9 The mass of the electron is $m_e = 0.9109 \times 10^{-27}$ g, and its energy equivalent is $m_e c^2 = 0.5110$ MeV. Show that the energy equivalent of the Planck mass is $m_{\rm P}c^2 = 1.221 \times 10^{19}$ GeV (1 GeV $= 10^9$ eV). This energy is called the Planck energy.

3.7 Gravitational potentials

We want to learn what happens with the gravitational constant G when we attempt to describe gravitation in spacetimes of other dimensionalities. To find out, we will examine gravitational potentials in Newtonian gravity. In this section we obtain the equation that relates the gravitational potential to the mass distribution in a spacetime with arbitrary but fixed number of spatial dimensions. In doing so we will learn how to define the relevant gravitational constant. This result will be used in the following section to define, in any dimension, the Planck length in terms of the appropriate gravitational constant.

We introduce a gravity field $\vec{g}$ with units of force per unit mass. The definition is similar to that of an electric field in terms of the force on a test particle: the force on a given test mass m at a point where the gravity field is $\vec{g}$ is given by $m\vec{g}$. We set $\vec{g}$ equal to minus the gradient of a gravitational potential V_g:

$$\vec{g} = -\nabla V_g. \tag{3.96}$$

We will take this equation to be true in all dimensions. Equation (3.96) has content: if you move a particle along a closed loop in a static gravitational field, the net work that you do against the gravitational field is zero.

Quick calculation 3.10 Prove the above statement.

What are the units for the gravitational potential? Equation (3.96) gives

$$[\,\vec{g}\,] = \frac{[\text{Force}]}{M} = \frac{[V_g]}{L} \longrightarrow [V_g] = \frac{[\text{Energy}]}{M}. \tag{3.97}$$

The gravitational potential has units of energy per unit mass in *any* dimension. The gravitational potential $V_g^{(4)}$ of a point mass in four dimensions is

$$V_g^{(4)} = -\frac{GM}{r}. \tag{3.98}$$

We can use the electromagnetic analogy to find the equation satisfied by the gravitational potential. In electromagnetism, we found an equation for the electrostatic potential which holds in any dimension. This is equation (3.76):

$$\nabla^2 \Phi = -\rho \,. \tag{3.99}$$

The four-dimensional scalar potential for a point charge q is

$$\Phi^{(4)} = \frac{q}{4\pi r}, \tag{3.100}$$

and it satisfies (3.99) where ρ is the charge density for the point charge. It follows by analogy that the four-dimensional gravitational potential in (3.98) satisfies

$$\nabla^2 V_g^{(4)} = 4\pi G \rho_m, \tag{3.101}$$

where ρ_m is the matter density. While this equation is correct in four dimensions, a small modification is needed for other dimensions. Note that the left-hand side has the same units in any number of dimensions: the units of V_g are always the same, and the Laplacian always divides by length-squared. The right-hand side must also have the same units in any number of dimensions. Since ρ_m is mass density, it has different units in different dimensions, and, as a consequence, the units of G must change when the dimensions change. We therefore rewrite the above equation more precisely as

$$\nabla^2 V_g^{(D)} = 4\pi G^{(D)} \rho_m, \tag{3.102}$$

when working in D-dimensional spacetime. The superscripts shown in parentheses denote the dimensionality of *spacetime*. In particular, we identify $G^{(4)}$ as the four-dimensional Newton constant G. In general, we will use D to denote the dimensionality of spacetime, and d to denote the number of spatial dimensions. Clearly, $D = d + 1$.

Equation (3.102) defines Newtonian gravitation in an arbitrary number of dimensions. Just as the electric field of a point charge does, the gravitational field of a point mass falls off like $1/r^{d-1}$ in a world with d spatial dimensions. As a result, the force between two point masses separated by a distance r falls off like $1/r^{d-1}$. For three spatial dimensions, this is the familiar inverse-squared dependence of the gravitational force. If $D = 6$ (a world with two extra dimensions) the gravitational force falls off like $1/r^4$.

3.8 The Planck length in various dimensions

We define the Planck length in any dimension just as we did in four dimensions: the Planck length is the unique length built using *only* powers of the gravitational constant $G^{(D)}$, c, and $\hbar$. To compute the Planck length we must determine the units of $G^{(D)}$. This is easily done if we recall that the units of $G^{(D)}\rho_m$ (the right-hand side of (3.102)) are the same in all dimensions.

Comparing the cases of five and four dimensions, for example,

$$[G^{(5)}]\,\frac{M}{L^4} = [G]\,\frac{M}{L^3} \;\longrightarrow\; [G^{(5)}] = L\,[G]. \tag{3.103}$$

The units of $G^{(5)}$ carry one more factor of length than the units of G. We use (3.90) to read the units of G in terms of units of length and units of c and $\hbar$:

$$[G] = \frac{[c]^3\, L^2}{[\hbar]}. \tag{3.104}$$

Equation (3.103) then gives

$$[G^{(5)}] = \frac{[c]^3\, L^3}{[\hbar]}. \tag{3.105}$$

Since the Planck length is constructed uniquely from the gravitational constant, c, and $\hbar$, we can remove the brackets in the above equation and replace L by the five-dimensional Planck length $\ell_{\rm P}^{(5)}$:

$$\left(\ell_{\rm P}^{(5)}\right)^3 = \frac{\hbar G^{(5)}}{c^3}. \tag{3.106}$$

Reintroducing the four-dimensional Planck length:

$$\left(\ell_{\rm P}^{(5)}\right)^3 = \left(\frac{\hbar G}{c^3}\right)\frac{G^{(5)}}{G} \longrightarrow \left(\ell_{\rm P}^{(5)}\right)^3 = (\ell_{\rm P})^2\frac{G^{(5)}}{G}. \tag{3.107}$$

Since they do not have the same units, the gravitational constants in four and five dimensions cannot be compared directly. Planck lengths, however, can be compared. If the Planck length is the same in four and in five dimensions, then $G^{(5)}/G = \ell_{\rm P}$; the gravitational constants differ by one factor of the common Planck length.

It is not hard to generalize the above equations to D spacetime dimensions:

Quick calculation 3.11 Show that (3.106) and (3.107) are replaced by

$$\left(\ell_{\rm P}^{(D)}\right)^{D-2} = \frac{\hbar G^{(D)}}{c^3} = (\ell_{\rm P})^2\frac{G^{(D)}}{G}. \tag{3.108}$$

3.9 Gravitational constants and compactification

If string theory is correct, our world is really higher-dimensional. The fundamental gravity theory is then defined in the higher-dimensional world, with some value for the higher-dimensional Planck length. Since we observe only four dimensions, the additional dimensions may be curled up to form a compact space with small volume. We can then ask: what is the effective value of the four-dimensional Planck length? As we shall show here, the effective four-dimensional Planck length depends on the volume of the extra dimensions, as well as on the value of the higher-dimensional Planck length.

These observations raise the possibility that the Planck length in the effectively four-dimensional world – the famous number equal to about 10^{-33} cm – may not coincide with

the fundamental Planck length in the original higher-dimensional theory. Is it possible that the fundamental Planck length is much bigger than the familiar, four-dimensional one? We will answer this question in the following section. In this section we will work out the effect of compactification on gravitational constants.

How do we calculate the gravitational constant in four dimensions if we are given the gravitational constant in five? First, we recognize the need to curl up one spatial dimension, otherwise there is no effectively four-dimensional spacetime. As we will see, the size of the extra dimension enters the relationship between the gravitational constants. To explore these questions precisely, consider a five-dimensional spacetime where one dimension forms a small circle of radius R. We are given $G^{(5)}$ and we would like to calculate $G^{(4)}$.

Let (x^1, x^2, x^3) denote three spatial dimensions of infinite extent, and x^4 denote a compactified dimension of circumference $2\pi R$ (Figure 3.2). We place a uniform ring of total mass M all around the circle at $x^1 = x^2 = x^3 = 0$. This is a mass distribution which is constant along the x^4 dimension. We are interested in the gravitational potential $V_g^{(5)}$ that emerges from such a mass distribution. Alternatively, we could have placed a point mass at some fixed x^4, but this makes the calculations more involved (Problem 3.10). In the present case the gravitational potential $V_g^{(5)}$ does *not* depend on x^4. The total mass M can be written as

$$\text{total mass} = M = 2\pi Rm, \tag{3.109}$$

where m is the mass per unit length.

What is the mass density in the five-dimensional world? It is only nonzero at $x^1 = x^2 = x^3 = 0$. To represent such a mass density we use delta functions. Recall that the delta function $\delta(x)$ can be viewed as a singular function whose value is zero except for $x = 0$ and such that the integral $\int_{-\infty}^{\infty} dx\delta(x) = 1$. This integral implies that if x has units of length, then $\delta(x)$ has units of inverse length. Since the five-dimensional mass density is concentrated at $x^1 = x^2 = x^3 = 0$, it is reasonable to include in its formula the product $\delta(x^1)\delta(x^2)\delta(x^3)$ of three delta functions. We claim that

$$\rho^{(5)} = m\,\delta(x^1)\delta(x^2)\delta(x^3). \tag{3.110}$$

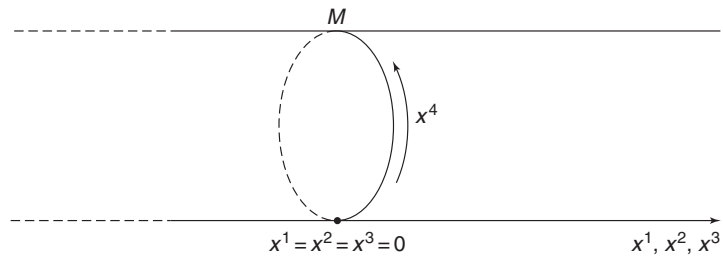

Fig. 3.2 A world with four space dimensions, one of which, x^4, is compactified into a circle of radius R. A ring of total mass M wraps around this compact dimension.

We first check the units. The mass density $\rho^{(5)}$ must have units of M/L^4. This works out since m has units of mass per unit length, and the three delta functions supply an additional factor of L^{-3}. The ansatz in (3.110) could still be off by a constant dimensionless factor, a factor of two, for example. As a final check, we integrate $\rho^{(5)}$ over all space. The result should be the total mass:

$$\begin{aligned}\int_{-\infty}^{\infty} dx^1 dx^2 dx^3 \int_0^{2\pi R} dx^4 \rho^{(5)} \\ = m \int_{-\infty}^{\infty} dx^1 \delta(x^1) \int_{-\infty}^{\infty} dx^2 \delta(x^2) \int_{-\infty}^{\infty} dx^3 \delta(x^3) \int_0^{2\pi R} dx^4 \\ = m\, 2\pi R. \end{aligned} \tag{3.111}$$

This is indeed the total mass on account of (3.109). For the effectively four-dimensional observer the mass is point-like, and it is located at $x^1 = x^2 = x^3 = 0$. So this observer writes

$$\rho^{(4)} = M\delta(x^1)\delta(x^2)\delta(x^3). \tag{3.112}$$

Note the relation

$$\rho^{(5)} = \frac{1}{2\pi R}\,\rho^{(4)}. \tag{3.113}$$

Let us now use this information in the equations for the gravitational potential. Using the five-dimensional version of (3.102) and (3.113), we find

$$\nabla^2 V_g^{(5)}(x^1, x^2, x^3) = 4\pi G^{(5)} \rho^{(5)} = 4\pi \frac{G^{(5)}}{2\pi R}\,\rho^{(4)}. \tag{3.114}$$

As we have noted before, $V_g^{(5)}$ is independent of x^4 so the Laplacian above is actually the four-dimensional one. Since the effective four-dimensional gravitational potential is $V_g^{(5)}$, the above equation takes the form of the gravitational equation in four dimensions, where the constant in between the 4π and the $\rho^{(4)}$ is the four-dimensional gravitational constant. We have therefore shown that

$$G = \frac{G^{(5)}}{2\pi R} \longrightarrow \boxed{\frac{G^{(5)}}{G} = 2\pi R \equiv \ell_C,} \tag{3.115}$$

where ℓ_C is the length of the extra compact dimension. This is what we were seeking: a relationship between the strength of the gravitational constants in terms of the size of the extra dimension.

The generalization of (3.115) to the case where there is more than one extra dimension is straightforward. One finds that

$$\frac{G^{(D)}}{G} = (\ell_C)^{D-4}, \tag{3.116}$$

where ℓ_C is the common length of each of the extra dimensions. When the various dimensions are curled up into circles of different lengths, the above right-hand side must be

replaced by the product of the various lengths. This product is, in fact, the volume V_C of the extra dimensions, so that

$$\frac{G^{(D)}}{G} = V_C. \tag{3.117}$$

3.10 Large extra dimensions

We are now done with all the groundwork. In Section 3.8 we found the relation between the Planck length and the gravitational constant in any dimension. In Section 3.9 we determined how gravitational constants are related upon compactification. We are ready to find out how the fundamental Planck length in a higher-dimensional theory with compactification is related to the Planck length in the effectively four-dimensional theory.

To begin with, consider a five-dimensional world with Planck length $\ell_{\rm P}^{(5)}$ and a single spatial coordinate curled up into a circle of circumference ℓ_C. How are these lengths related to $\ell_{\rm P}$? From (3.107) and (3.115) we find that

$$(\ell_{\rm P}^{(5)})^3 = (\ell_{\rm P})^2 \, \frac{G^{(5)}}{G} = (\ell_{\rm P})^2 \, \ell_C. \tag{3.118}$$

Solving for ℓ_C, we get

$$\ell_C = \frac{(\ell_{\rm P}^{(5)})^3}{(\ell_{\rm P})^2}. \tag{3.119}$$

This relation enables us to explore the possibility that the world is actually five-dimensional with a fundamental Planck length $\ell_{\rm P}^{(5)}$ that is much larger than 10^{-33} cm. Of course, we must have $\ell_{\rm P} \sim 10^{-33}$ cm. After all, this is the four-dimensional Planck length, whose value is given in (3.90).

Present-day accelerators explore physics down to distances of the order of 10^{-16} cm. If this distance, or a somewhat smaller one, is the fundamental length scale, we may choose $\ell_{\rm P}^{(5)} \sim 10^{-18}$ cm. What would ℓ_C have to be? With $\ell_{\rm P}^{(5)} \sim 10^{-18}$ cm and $\ell_{\rm P} \sim 10^{-33}$ cm, equation (3.119) gives $\ell_C \sim 10^{12}$ cm $\sim 10^7$ km. This is more than twenty times the distance from the earth to the moon. Such a large extra dimension would have been detected a long time ago.

Having failed to produce a realistic scenario in five dimensions, let us try in *six* spacetime dimensions. For arbitrary D, equations (3.108) and (3.116) give

$$\left(\ell_{\rm P}^{(D)}\right)^{D-2} = (\ell_{\rm P})^2 \frac{G^{(D)}}{G} = (\ell_{\rm P})^2 (\ell_C)^{D-4}. \tag{3.120}$$

Solving for ℓ_C we find

$$\ell_C = \ell_{\rm P}^{(D)} \left(\frac{\ell_{\rm P}^{(D)}}{\ell_{\rm P}}\right)^{\frac{2}{D-4}}. \tag{3.121}$$

For $D = 6$ and $\ell_{\mathrm{P}}^{(6)} \sim 10^{-18}$ cm this formula gives

$$\ell_C = \frac{(\ell_{\mathrm{P}}^{(6)})^2}{\ell_{\mathrm{P}}} \sim 10^{-3} \text{ cm}. \tag{3.122}$$

This is a lot more interesting! A convenient unit here is the micron $\mu\text{m} = 10^{-6}$ m. One-tenth of a millimeter is 100 μm. We found $\ell_C \sim 10\,\mu$m. Could there be extra dimensions ten microns long? You might think that this is still too big, since even microscopes probe smaller distances. Moreover, as we indicated before, accelerators probe distances of the order of 10^{-16} cm. Surprisingly, it is possible that "large extra dimensions" exist and that we have not observed them yet.

The existence of additional dimensions may be confirmed by testing the force law which gives the gravitational attraction between two masses. For distances much larger than the compactification scale ℓ_C the world is effectively four-dimensional, so the dependence of the force between two masses on their separation must follow accurately Newton's inverse-squared law. On the other hand, for distances smaller than ℓ_C, the world is effectively higher-dimensional, and the force law will change. A force between two masses that goes like $1/r^4$, where r is the separation, is consistent with the existence of two compact extra dimensions.

It turns out to be very difficult to test gravity at small distances; the force of gravity is extremely weak and spurious electrical forces must be cancelled very precisely. Motivated mainly by the possible existence of large extra dimensions, physicists set out to test the inverse-squared law at distances smaller than one millimeter. The tabletop experiments use a torsion pendulum detector or, alternatively, a micromachined cantilever with a test mass at the free end. As of 2007, experiments have found no departure from the inverse-squared law down to distances of about fifty microns. This means that extra dimensions, if they exist, must be smaller than this distance. Compact dimensions the size of ten microns, as we found in (3.122), are still consistent with experiment.

You might ask: what about forces other than gravity? Electromagnetism has been tested to much smaller distances, and we know that the electric force obeys an inverse-squared law very accurately. Rutherford scattering of alpha particles off nuclei, for example, confirms that the inverse-squared law holds down to 10^{-11} cm. Since the separation dependence of the electric force would change at distances smaller than the size of the extra dimensions, this seems to rule out large extra dimensions. The possibility of large extra dimensions, however, survives in string theory, where our spatial world could be a three-dimensional hyperplane transverse to the extra dimensions. This hyperplane is called a D3-brane. A D3-brane is a D-brane with three spatial dimensions.

Open strings have the remarkable property that their endpoints must remain attached to the D-branes. In many phenomenological models built in string theory, it is the fluctuations of open strings that give rise to the familiar leptons, quarks, and gauge fields, including the Maxwell gauge field. It follows that these fields are bound to the D3-brane and do *not* feel the extra dimensions. If the Maxwell field lives on the D-brane, the electric field lines of a charge remain on the D-brane and do not go off into the extra dimensions. The force law is not changed at any distance scale. Closed strings are not bound by D-branes, and therefore gravity, which arises from closed strings, *is* affected by the extra dimensions.

Although the Planck length ℓ_P is an important length scale in four dimensions, if there are large extra dimensions, the truly fundamental Planck length would be much bigger than the effective four-dimensional one. The possibility of large extra dimensions is slightly unnatural – why should the extra dimensions be much larger than the fundamental length scale? This is not a new problem, however, but rather the problem of a large hierarchy in another guise. We noted earlier that particle physics faces a puzzling hierarchy between the Planck mass and the masses of elementary particles. In the large-extra-dimensions scenario, the hierarchy is postulated to arise from extra dimensions that are much larger than the fundamental length scale. At any rate, the truly exciting fact is that present experimental constraints do not rule out large extra dimensions. The discovery of extra dimensions would be revolutionary.

It is possible, of course, that extra dimensions are much larger than $\ell_P \sim 10^{-33}$ cm but are still quite small. If spacetime is ten dimensional, setting $D = 10$ and $\ell_P^{(10)} \sim 10^{-18}$ cm in (3.121) gives $\ell_C \sim 10^{-13}$ cm, a distance far too small to be probed with tabletop gravitational experiments. Extra dimensions this size or smaller must be searched for using particle accelerators.

Problems

Problem 3.1 Lorentz covariance for motion in electromagnetic fields.[†]

The Lorentz force equation (3.5) can be written relativistically as

$$\frac{dp_\mu}{ds} = \frac{q}{c} F_{\mu\nu} \frac{dx^\nu}{ds}, \tag{1}$$

where p_μ is the four-momentum. Check explicitly that this equation reproduces (3.5) when μ is a spatial index. What does (1) give when $\mu = 0$? Does it make sense? Is (1) a gauge invariant equation?

Problem 3.2 Maxwell equations in four dimensions.

(a) Show explicitly that the source-free Maxwell equations emerge from $T_{\mu\lambda\nu} = 0$.
(b) Show explicitly that the Maxwell equations with sources emerge from (3.34).

Problem 3.3 Electromagnetism in three dimensions.

(a) Find the reduced Maxwell equations in three dimensions by starting with Maxwell's equations and the force law in four dimensions, using the ansatz (3.11), and assuming that no field can depend on the z direction.
(b) Repeat the analysis of three-dimensional electromagnetism starting with the Lorentz covariant formulation. Take $A^\mu = (\Phi, A^1, A^2)$, examine $F_{\mu\nu}$, the Maxwell equations (3.34), and the relativistic form of the force law derived in Problem 3.1.

Problem 3.4 Electric fields and potentials of point charges.

(a) Show that for time-independent fields, the Maxwell equation $T_{0ij} = 0$ implies that $\partial_i E_j - \partial_j E_i = 0$. Explain why this condition is satisfied by the ansatz $\vec{E} = -\nabla\Phi$.

(b) Show that with d spatial dimensions, the potential Φ due to a point charge q is given by

$$\Phi(r) = \frac{\Gamma\left(\frac{d}{2}-1\right)}{4\pi^{d/2}} \frac{q}{r^{d-2}}.$$

Problem 3.5 Calculating the divergence in higher dimensions.

Let $\vec{f} = f(r)\,\hat{\mathbf{r}}$ be a vector function in $\mathbb{R}^d$. Here $\hat{\mathbf{r}}$ is a unit radial vector, and r is the radial distance to the origin. Derive a formula for $\nabla \cdot \vec{f}$ by applying the divergence theorem to a spherical shell of radius r and width dr. Check that for $d = 3$ your answer reduces to $\nabla \cdot \vec{f} = f'(r) + \frac{2}{r} f(r)$.

Problem 3.6 Analytic continuation for gamma functions.[†]

Consider the definition of the gamma function for complex arguments z whose real part is positive:

$$\Gamma(z) = \int_0^\infty dt\, e^{-t} t^{z-1}, \quad \Re(z) > 0.$$

Use this equation to show that for $\Re(z) > 0$

$$\Gamma(z) = \int_0^1 dt\, t^{z-1}\Big(e^{-t} - \sum_{n=0}^{N} \frac{(-t)^n}{n!}\Big) + \sum_{n=0}^{N} \frac{(-1)^n}{n!} \frac{1}{z+n} + \int_1^\infty dt\, e^{-t} t^{z-1}.$$

Explain why the above right-hand side is well defined for $\Re(z) > -N - 1$. It follows that this right-hand side provides the analytic continuation of $\Gamma(z)$ for $\Re(z) > -N - 1$. Conclude that the gamma function has poles at $0, -1, -2, \ldots$, and give the value of the residue at $z = -n$ (with n a positive integer).

Problem 3.7 Simple quantum gravity effects are small.[†]

(a) What would be the "gravitational" Bohr radius for a hydrogen atom if the attraction binding the electron to the proton was gravitational? The standard Bohr radius is $a_0 = \frac{\hbar^2}{me^2} \simeq 5.29 \times 10^{-9}$ cm.

(b) In "units" where G, c, and $\hbar$ are set equal to one, the temperature of a black hole is given by $kT = \frac{1}{8\pi M}$. Insert back the factors of G, c, and $\hbar$ into this formula. Evaluate the temperature of a black hole of a million solar masses. What is the mass of a black hole whose temperature is room temperature?

Problem 3.8 Vacuum energy and an associated length scale.

Observations indicate that the expansion of the universe is currently accelerating possibly due to a vacuum energy density. The mass density associated with this energy is approximately $\rho_{\text{vac}} = 7.7 \times 10^{-27}$ kg/m^3. Some physicists try to understand the acceleration of the universe by introducing modifications to gravity. It is then useful to know what length scales could be important. If one assumes that the only relevant parameters are ρ_{vac}, $\hbar$, and c, one can construct a length parameter ℓ_{vac} by multiplying powers:

$$\ell_{\text{vac}} = \rho_{\text{vac}}^{\alpha}\, \hbar^{\beta}\, c^{\gamma}\ .$$

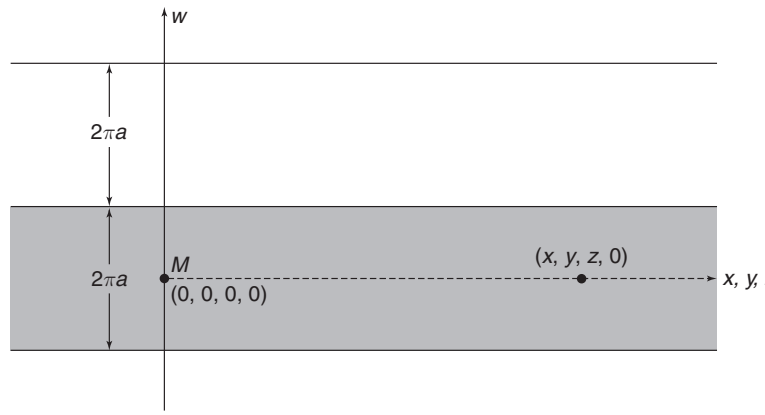

Fig. 3.3 Problem 3.10: a point mass M in a five-dimensional spacetime with one compact dimension.

What must be the values of α, β, and γ in the above equation? What is the numerical value of $\ell_{\rm vac}$? Express your answer in μm, where 1 μm = 10^{-6} m.

Problem 3.9 Planetary motion in four and higher dimensions.

Consider the motion of planets in planar circular orbits around heavy stars in our four-dimensional spacetime and in spacetimes with additional spatial dimensions. We wish to study the stability of these orbits under perturbations that keep them planar. Such a perturbation would arise, for example, if a meteorite moving on the plane of the orbit hits the planet and changes its angular momentum.

Show that while planetary circular orbits in our four-dimensional world are stable under such perturbations, they are not so in five or higher dimensions. [Hint: you may find it useful to use the effective potential for motion in a central force field.]

Problem 3.10 Gravitational field of a point mass in a compactified five-dimensional world.

Consider a five-dimensional spacetime with space coordinates (x, y, z, w) *not* yet compactified. A point mass M is located at the origin $(x, y, z, w) = (0, 0, 0, 0)$.

(a) Find the gravitational potential $V_g^{(5)}(r)$. Write your answer in terms of M, $G^{(5)}$, and $r = (x^2 + y^2 + z^2 + w^2)^{1/2}$. [Hint: use $\nabla^2 V_g^{(5)} = 4\pi G^{(5)} \rho_m$ and the divergence theorem.]

Now let w become a circle with radius a while keeping the mass fixed, as shown in Figure 3.3.

(b) Write an exact expression for the gravitational potential $V_g^{(5)}(x, y, z, 0)$. This potential is a function of $R \equiv (x^2 + y^2 + z^2)^{1/2}$ and can be written as an infinite sum.

(c) Show that for $R \gg a$ the gravitational potential takes the form of a four-dimensional gravitational potential, with Newton's constant $G^{(4)}$ given in terms of $G^{(5)}$ as in (3.115). [Hint: turn the infinite sum into an integral.]

These results confirm both the relation between the four- and five-dimensional Newton constants in a compactification and the emergence of a four-dimensional potential at distances large compared to the size of the compact dimension.

Problem 3.11 Exact answer for the gravitational potential.

The infinite sum in Problem 3.10 can be evaluated exactly using the identity

$$\sum_{n=-\infty}^{\infty} \frac{1}{1+(\pi n x)^2} = \frac{1}{x} \coth\Big(\frac{1}{x}\Big).$$

(a) Find an exact closed-form expression for the gravitational potential $V_g^{(5)}(x, y, z, 0)$ in the compactified theory.
(b) Expand this answer to calculate the leading *correction* to the gravitational potential in the limit when $R \gg a$. For what value of R/a is the correction of order 1%?
(c) Use the exact answer in (a) to expand the potential when $R \ll a$. Give the first two terms in the expansion. Do you recognize the leading term?

4 Nonrelativistic strings

A full appreciation for the subtleties of relativistic strings requires an understanding of the basic physics of nonrelativistic strings. These strings have mass and tension. They can vibrate both transversely and longitudinally. We study the equations of motion for nonrelativistic strings and develop the Lagrangian approach to their dynamics.

4.1 Equations of motion for transverse oscillations

We will begin our study of strings with a look at the transverse fluctuations of a stretched string. The direction along the string is called the longitudinal direction, and the directions orthogonal to the string are called the transverse directions. We consider, for notational simplicity, the case when there is only one transverse direction – the generalization to additional transverse directions is straightforward.

Working in the (x, y) plane, let the classical nonrelativistic string have its endpoints fixed at $(0, 0)$, and $(a, 0)$. In the static configuration the string is stretched along the x axis between these two points. In a transverse oscillation, the x-coordinate of any point on the string does not change in time. The transverse displacement of a point is given by its y coordinate. The x direction is longitudinal, and the y direction is transverse. To describe the classical mechanics of a homogeneous string, we need two pieces of information: the tension T_0 and the mass per unit length μ_0. The total mass of the string is then $M = \mu_0 a$.

Let us look briefly at the units. Tension has units of force, so

$$[T_0] = [\text{Force}] = \frac{[\text{Energy}]}{L}. \tag{4.1}$$

If you stretch a string an infinitesimal amount dx, its tension remains approximately constant through the stretching, and the change in energy equals the work done $T_0 dx$. The total mass of the string does not change. If we were considering relativistic strings, however, a static string with more energy would have a larger rest mass. Using (4.1), noting that energy has units of mass times velocity squared, and that μ_0 has units of mass per unit length, we have

$$[T_0] = \frac{M}{L}[v]^2 = [\mu_0][v]^2. \tag{4.2}$$

For a nonrelativistic string, both T_0 and μ_0 are adjustable parameters, and the velocity on the right-hand side above will turn out to be the velocity of transverse waves. The above

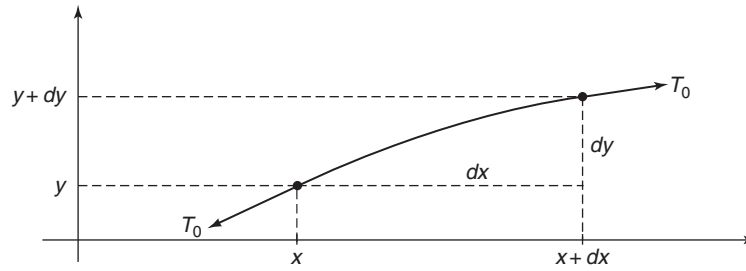

Fig. 4.1 **A short piece of a classical nonrelativistic string vibrating transversely. With different slopes at the two endpoints there is a net vertical force.**

equation suggests that the string tension T_0 and the linear mass density μ_0 in a relativistic string might be related by $T_0 = \mu_0 c^2$, since c is the canonical velocity in relativity. We will see in Chapter 6 that this is indeed the correct relation for a relativistic string.

Returning to our classical nonrelativistic string, let us figure out the equation of motion. Consider a small portion of the static string that extends from x to $x + dx$, with $y = 0$. This piece is shown in transverse oscillation in Figure 4.1. At time t, the transverse displacement of the string is $y(t, x)$ at x and $y(t, x + dx)$ at $x + dx$. We will assume that the oscillations are small, and by this we will mean that at all times

$$\left|\frac{\partial y}{\partial x}\right| \ll 1, \tag{4.3}$$

at any point on the string. This guarantees that the transverse displacement of the string is small compared to the length of the string. The length of the string changes little, and we can assume that the tension T_0 is unchanged.

The slope of the string is a bit different at the points x and $x + dx$. This change of slope means that the string tension changes direction and the portion of string under consideration feels a net force. For transverse oscillations we need only calculate the net vertical force; the net horizontal force is negligible (Problem 4.1). The vertical force at $(x + dx, y + dy)$ is given accurately by T_0 times $\partial y/\partial x$ evaluated at $x + dx$ and is pointing up; similarly, the vertical force at (x, y) is T_0 times $\partial y/\partial x$ evaluated at x and is pointing down. Therefore the net vertical force dF_v is

$$dF_v = T_0 \frac{\partial y}{\partial x}\bigg|_{x+dx} - T_0 \frac{\partial y}{\partial x}\bigg|_{x} \simeq T_0 \frac{\partial^2 y}{\partial x^2} dx. \tag{4.4}$$

The mass dm of this piece of string, originally stretched from x to $x + dx$, is given by the mass density μ_0 times dx. By Newton's law, the net vertical force equals mass times vertical acceleration. So we can simply write

$$T_0 \frac{\partial^2 y}{\partial x^2} dx = (\mu_0 dx)\, \frac{\partial^2 y}{\partial t^2}. \tag{4.5}$$

We cancel dx on each side and rearrange terms to get

$$\frac{\partial^2 y}{\partial x^2} - \frac{\mu_0}{T_0}\frac{\partial^2 y}{\partial t^2} = 0\,. \tag{4.6}$$

This is just a wave equation! Recall that for the wave equation

$$\frac{\partial^2 y}{\partial x^2} - \frac{1}{v_0^2}\frac{\partial^2 y}{\partial t^2} = 0, \tag{4.7}$$

the parameter v_0 is the velocity of the waves. Thus for the transverse waves on our stretched string, the velocity v_0 of the waves is

$$v_0 = \sqrt{T_0/\mu_0}. \tag{4.8}$$

The higher the tension or the lighter the string, the faster the waves move.

4.2 Boundary conditions and initial conditions

Since equation (4.6) is a partial differential equation involving space and time derivatives, in order to fix solutions we must in general apply both boundary conditions and initial conditions. Boundary conditions (B.C.) constrain the solution at the boundary of the system, and initial conditions constrain the solution at a given starting time. The most common types of boundary conditions are Dirichlet and Neumann boundary conditions.

For our string, Dirichlet boundary conditions specify the positions of the string endpoints. For example, if we attach each end of the string to a wall (Figure 4.2, left), we are imposing the Dirichlet boundary conditions

$$y(t, x = 0) = y(t, x = a) = 0, \quad \text{Dirichlet boundary conditions.} \tag{4.9}$$

Alternatively, if we attach a massless loop to each end of the string and the loops are allowed to slide along two frictionless poles, we are imposing Neumann boundary conditions. For our string, Neumann boundary conditions specify the values of the derivative $\partial y/\partial x$ at the endpoints. Since the loops are massless and the poles are frictionless, the derivative $\partial y/\partial x$ must vanish at the poles $x = 0, a$ (Figure 4.2, right). If this were not the case, then the slope of the string at a pole would be nonzero, and a component of the string tension would accelerate the rings in the y direction. Since each ring is massless,

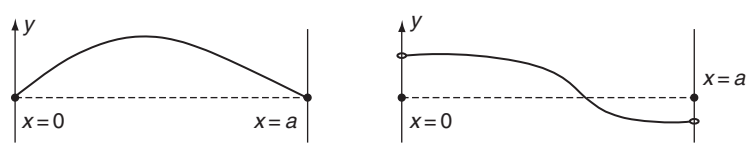

Fig. 4.2 **Left: string with Dirichlet boundary conditions at the endpoints. Right: string with Neumann boundary conditions at the endpoints.**

their acceleration would be infinite. This is not possible, so, in effect, we are imposing the Neumann boundary conditions

$$\frac{\partial y}{\partial x}(t, x = 0) = \frac{\partial y}{\partial x}(t, x = a) = 0, \quad \text{Neumann boundary conditions.} \tag{4.10}$$

These Neumann boundary conditions apply to strings whose endpoints are free to move along the y direction.

Let us see how we can solve the wave equation for a particular set of initial conditions. The general solution of equation (4.6) is of the form

$$y(t, x) = h_+(x - v_0 t) + h_-(x + v_0 t), \tag{4.11}$$

where h_+ and h_- are arbitrary functions of a *single* variable. This solution represents a superposition of two waves, h_+ moving to the right and h_- moving to the left. Suppose that the initial values of y and $\partial y/\partial t$ are known at time $t = 0$. Using equation (4.11) we see that this information yields the equations

$$y(0, x) = h_+(x) + h_-(x), \tag{4.12}$$

$$\frac{\partial y}{\partial t}(0, x) = -v_0 h'_+(x) + v_0 h'_-(x), \tag{4.13}$$

where the left-hand sides are known functions, and the primes denote derivatives with respect to arguments. Using (4.12) we can solve for h_- in terms of h_+. Substituting into (4.13), we get a first-order ordinary differential equation for h_+. Once we have solved for h_+ (using appropriate boundary conditions), we can use (4.12) again, this time to find the explicit form of h_-. With h_+ and h_- known, the full solution of the equations of motion is given by (4.11).

4.3 Frequencies of transverse oscillation

Suppose that we have a string where each point is oscillating in the y direction sinusoidally and in phase. This means that $y(t, x)$ is of the form

$$y(t, x) = y(x) \sin(\omega t + \phi), \tag{4.14}$$

where ω is the angular frequency of oscillation and ϕ is the constant common phase. Our aim is to find the allowed frequencies of oscillation. Substituting (4.14) into (4.6) and cancelling the common time dependence, we find

$$\frac{d^2 y(x)}{dx^2} + \omega^2 \frac{\mu_0}{T_0} y(x) = 0. \tag{4.15}$$

This is an ordinary second-order differential equation for the profile $y(x)$ of the oscillations. The allowed frequencies are selected by this equation, together with the boundary conditions. Since ω, μ_0, and T_0 are constants, the differential equation is solved in terms of

trigonometric functions. With Dirichlet boundary conditions (4.9) we have the nontrivial solutions

$$y_n(x) = A_n \sin\left(\frac{n\pi x}{a}\right), \quad n = 1, 2, \ldots, \tag{4.16}$$

where A_n is an arbitrary constant. The value $n = 0$ is not included above because it represents a motionless string. Plugging $y_n(x)$ into (4.15), we find the allowed frequencies ω_n:

$$\omega_n = \sqrt{\frac{T_0}{\mu_0}}\left(\frac{n\pi}{a}\right), \quad n = 1, 2, \ldots. \tag{4.17}$$

These are the frequencies of oscillation for a Dirichlet string. The strings on a violin are Dirichlet strings. To tune a violin to the correct frequency one must adjust the string tension. The higher the tension is, the higher the pitch, as predicted by (4.17). For the case of Neumann boundary conditions (4.10), we obtain the spatial solutions

$$y_n(x) = A_n \cos\left(\frac{n\pi x}{a}\right), \quad n = 0, 1, 2, \ldots. \tag{4.18}$$

This time the $n = 0$ solution is a little less trivial: the string does not oscillate, but it is rigidly translated to $y(t, x) = A_0$. The oscillation frequencies, found by plugging (4.18) into (4.15), are the same as those in (4.17). Therefore, the oscillation frequencies are the same in the Neumann and Dirichlet problems. The Neumann case admits one extra solution not included in our oscillatory ansatz (4.14): the string can translate with constant velocity. Indeed, $y(t, x) = at + b$, with a and b arbitrary constants, satisfies both the boundary conditions and the original wave equation (4.7).

4.4 More general oscillating strings

Let us discuss briefly some problems that are closely related to the ones considered thus far. For example, we can take the mass density of the string to be a function $\mu(x)$ of position. The form (4.6) of the wave equation does not change since it is derived from local considerations: the examination of a little piece of string that can be chosen to be sufficiently small so that the mass density is approximately constant. We therefore get

$$\frac{\partial^2 y}{\partial x^2} - \frac{\mu(x)}{T_0}\frac{\partial^2 y}{\partial t^2} = 0. \tag{4.19}$$

For normal oscillations, we use the ansatz in (4.14) and find

$$\frac{d^2 y}{dx^2} + \frac{\mu(x)}{T_0}\omega^2 y(x) = 0. \tag{4.20}$$

This equation is no longer simple to solve, and it can only be studied in detail once the function $\mu(x)$ is specified. In Problems 4.3 and 4.7 you will consider some specific mass distributions, and you will explore a variational approach that gives an upper bound for the lowest oscillation frequency.

So far we have only considered strings that oscillate transversally. Strings also admit longitudinal oscillations, although the relativistic string does not. Imagine a string stretched along the x axis, and consider the infinitesimal segment which at equilibrium extends from x to $x + dx$. Suppose now that at time t the ends of this infinitesimal segment are longitudinally displaced from their equilibrium positions by distances $\eta(t, x)$ and $\eta(t, x + dx)$, respectively. If these two quantities are not the same, the piece of string is being compressed or stretched. An equation of motion can be obtained for this system, much as we did for transverse motion. It is not possible, however, to assume that the tension is constant throughout the string. For transverse oscillations the net force acting on a little piece of string arose from the different angles at which the same tension was applied on opposite ends of the piece. If the string always lies along the x axis then a net force can act on a segment only if the tension is different on its two ends. Therefore the waves on a longitudinally oscillating string are accompanied by tension waves (Problem 4.2).

4.5 A brief review of Lagrangian mechanics

The Lagrangian L of a system is defined by

$$L = T - V\,, \tag{4.21}$$

where T is the kinetic energy of the system and V is the potential energy of the system. For a point particle of mass m moving along the x axis under the influence of a time-independent potential $V(x)$, the nonrelativistic Lagrangian takes the form

$$L(t) = \frac{1}{2}m\,(\dot{x}(t))^2 - V(x(t))\,, \quad \dot{x}(t) \equiv \frac{dx(t)}{dt}. \tag{4.22}$$

We must emphasize that the above Lagrangian is implicitly a function of time, but it has no explicit time dependence. All the time dependence arises from the time dependence of the position $x(t)$. The action S is defined as

$$S = \int_{\mathcal{P}} L(t)dt, \tag{4.23}$$

where $\mathcal{P}$ is a path $x(t)$ between an initial position x_i at an initial time t_i, and a final position x_f at a final time $t_f > t_i$. One such path is shown in Figure 4.3.

The action is a *functional*. Whereas a function of a single variable takes one number – the argument – as input and gives another number as output, a functional takes a *function* as the input, and gives a number as output. Since a function is usually defined by its values at infinitely many points, we can think of a functional as a function of infinitely many variables. In our present application, the input for the action functional is the function $x(t)$ which determines the path $\mathcal{P}$. We can emphasize the argument of S by using the notation $S[x]$. Here $[x]$ represents the full function $x(t)$. It is potentially confusing to write $S[x(t)]$, since it suggests that S is ultimately a function of t, which it is not.

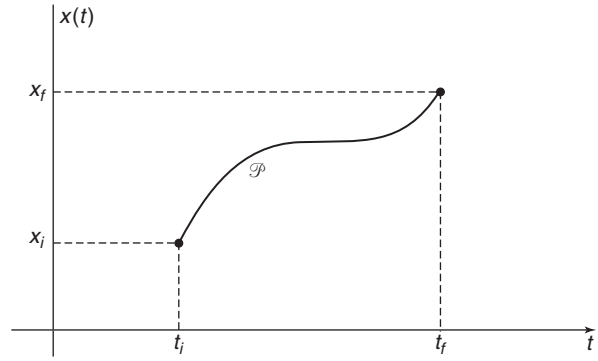

Fig. 4.3 A path $\mathcal{P}$ representing a possible one-dimensional motion $x(t)$ of a particle during the time interval $[t_i, t_f]$.

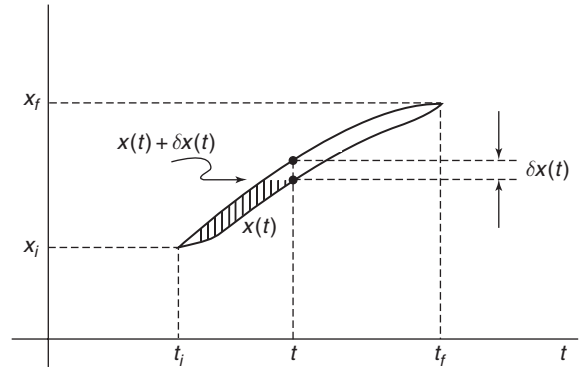

Fig. 4.4 A path $x(t)$ and its variation $x(t) + \delta x(t)$. This variation $\delta x(t)$ vanishes at $t = t_i$ and at $t = t_f$.

More explicitly, for any path $x(t)$, the action is given by

$$S[x] = \int_{t_i}^{t_f} \left\{ \frac{1}{2} m \left(\dot{x}(t)\right)^2 - V(x(t)) \right\} dt. \tag{4.24}$$

It is very important to emphasize that the action S can be calculated for any path $x(t)$ and not only for paths that represent physically realized motion. It is because S can be calculated for all paths that it is a very powerful tool to find the paths that can be physically realized.

Hamilton's principle states that the path $\mathcal{P}$ which a system actually takes is one for which the action S is stationary. More precisely, if this path $\mathcal{P}$ is varied infinitesimally, the action does not change to first order in the variation. In terms of the function $x(t)$ which specifies the path, the perturbed path takes the form $x(t) + \delta x(t)$, as shown in Figure 4.4. For any time t, the variation $\delta x(t)$ is the vertical distance between the original path and the varied path. As in the figure, we consider variations where the initial and final positions $x_i = x(t_i)$ and $x_f = x(t_f)$ are unchanged:

$$\delta x(t_i) = \delta x(t_f) = 0. \tag{4.25}$$

We now calculate the action $S[x+\delta x]$ for the perturbed path $x(t)+\delta x(t)$:

$$\begin{aligned} S[x+\delta x] &= \int_{t_i}^{t_f} \left\{ \frac{m}{2}\Big(\frac{d}{dt}(x(t)+\delta x(t))\Big)^2 - V\,(x(t)+\delta x(t)) \right\} dt \\ &= S[x] + \int_{t_i}^{t_f} \left\{ m\dot{x}(t)\frac{d}{dt}\delta x(t) - V'(x(t))\,\delta x(t) \right\} dt + \mathcal{O}((\delta x)^2). \end{aligned} \tag{4.26}$$

In passing to the last right-hand side we expanded V in a Taylor series about $x(t)$. The terms of order $(\delta x)^2$ and higher are unnecessary to determine whether or not the action is stationary. We have thus left them undetermined and indicated them by $\mathcal{O}((\delta x)^2)$. We can write the new action as $S+\delta S$, where δS is linear in δx. From the equation above we see that δS is given by

$$\delta S = \int_{t_i}^{t_f} \left\{ m\dot{x}(t)\frac{d}{dt}\delta x(t) - V'(x(t))\,\delta x(t) \right\} dt. \tag{4.27}$$

To find the equations of motion, the variation δS must be rewritten in the form $\delta S = \int dt\,\delta x(t)\{\ldots\}$. In particular, no derivatives must be acting on δx. This can be achieved using integration by parts:

$$\begin{aligned} \delta S &= \int_{t_i}^{t_f} \left\{ \frac{d}{dt}\big(m\dot{x}(t)\delta x(t)\big) - m\ddot{x}(t)\delta x(t) - V'(x(t))\,\delta x(t) \right\} dt \\ &= m\dot{x}(t_f)\delta x(t_f) - m\dot{x}(t_i)\delta x(t_i) + \int_{t_i}^{t_f} \delta x(t)\big(-m\ddot{x}(t) - V'(x(t))\big)\,dt. \end{aligned} \tag{4.28}$$

Making use of (4.25), the variation reduces to

$$\delta S = \int_{t_i}^{t_f} \delta x(t)\big(-m\ddot{x}(t) - V'(x(t))\big)\,dt. \tag{4.29}$$

The action is stationary if δS vanishes for every variation $\delta x(t)$. For this to happen, the factor multiplying $\delta x(t)$ in the integrand must vanish:

$$m\ddot{x}(t) = -V'(x(t)). \tag{4.30}$$

This is Newton's second law applied to the motion of a particle in a potential $V(x)$. We have recovered the expected equation of motion by requiring that the action be stationary under variations.

Suppose that we have determined the path that the particle takes while going from x_i to x_f. As we have seen, the action is then stationary under variations that vanish at the initial and final times. Is the action also stationary under variations that change the initial position at t_i or the final position at t_f? In general, the answer is no. This can be seen from equation (4.5). The integral term vanishes by assumption, but if $\delta x(t_f) \neq 0$, the first term on the right-hand side would not vanish unless $m\dot{x}(t_f)$, the final momentum of the particle, happens to vanish. The situation is analogous for $\delta x(t_i) \neq 0$.

Hamilton's principle states that the action is stationary about the classical solution. The classical solution does not always define a minimum of the action. It is possible to construct a simple example in which the classical solution is a saddle point of the action functional:

the action increases for some variations and decreases for others. See Problem 4.6 for the details.

4.6 The nonrelativistic string Lagrangian

Let us return now to our string with constant mass density μ_0, constant tension T_0, and ends located at $x = 0$ and $x = a$. The kinetic energy is simply the sum of the kinetic energies of all the infinitesimal segments that comprise the string. So it can be written as

$$T = \int_0^a \frac{1}{2}\,(\mu_0 dx)\left(\frac{\partial y}{\partial t}\right)^2. \tag{4.31}$$

The potential energy arises from the work which must be done to stretch the segments. Consider an infinitesimal portion of string which extends from $(x, 0)$ to $(x + dx, 0)$ when the string is in equilibrium. If the string element is momentarily stretched from (x, y) to $(x + dx, y + dy)$, as in Figure 4.1, then the change in length Δl of the infinitesimal segment is given by

$$\Delta l = \sqrt{(dx)^2 + (dy)^2} - dx = dx\left(\sqrt{1 + \left(\frac{\partial y}{\partial x}\right)^2} - 1\right) \simeq dx\,\frac{1}{2}\left(\frac{\partial y}{\partial x}\right)^2, \tag{4.32}$$

where we have used the small oscillation approximation (4.3) to discard higher-order terms in the expansion of the square root. Since the work done in stretching each infinitesimal segment is $T_0 \Delta l$, the total potential energy V is

$$V = \int_0^a \frac{1}{2} T_0 \left(\frac{\partial y}{\partial x}\right)^2 dx. \tag{4.33}$$

The Lagrangian for the string is given by $T - V$:

$$L(t) = \int_0^a \left[\frac{1}{2}\mu_0\left(\frac{\partial y}{\partial t}\right)^2 - \frac{1}{2}T_0\left(\frac{\partial y}{\partial x}\right)^2\right] dx \equiv \int_0^a \mathcal{L}\,dx\,, \tag{4.34}$$

where $\mathcal{L}$ is referred to as the *Lagrangian density*:

$$\mathcal{L}\left(\frac{\partial y}{\partial t}, \frac{\partial y}{\partial x}\right) = \frac{1}{2}\mu_0\left(\frac{\partial y}{\partial t}\right)^2 - \frac{1}{2}T_0\left(\frac{\partial y}{\partial x}\right)^2. \tag{4.35}$$

The action for our string is therefore

$$S = \int_{t_i}^{t_f} L(t)dt = \int_{t_i}^{t_f} dt \int_0^a dx \left[\frac{1}{2}\mu_0\left(\frac{\partial y}{\partial t}\right)^2 - \frac{1}{2}T_0\left(\frac{\partial y}{\partial x}\right)^2\right]. \tag{4.36}$$

In this action the "path" is the function $y(t, x)$, defined over the region of (t, x) space shown shaded in Figure 4.5.

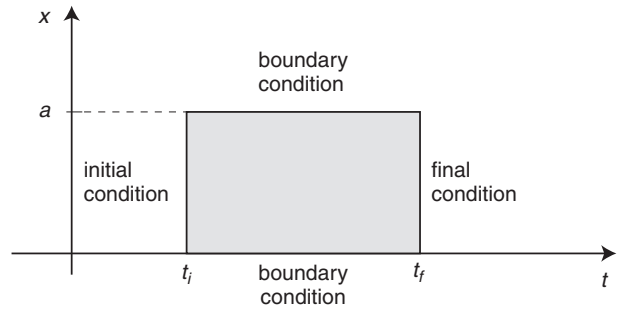

Fig. 4.5 **The string motion is defined by $y(t, x)$ over the domain $t \in [t_i, t_f]$, $x \in [0, a]$. Boundary conditions apply at $x = 0$ and $x = a$ for all $t \in [t_i, t_f]$. Initial and final conditions apply for $t = t_i$ and $t = t_f$, respectively, for all $x \in [0, a]$.**

To find the equations of motion, we must examine the variation of the action as we vary: $y(t, x) \to y(t, x) + \delta y(t, x)$. Performing the variation as before, we get

$$\delta S = \int_{t_i}^{t_f} dt \int_0^a dx \left[\mu_0 \frac{\partial y}{\partial t} \frac{\partial (\delta y)}{\partial t} - T_0 \frac{\partial y}{\partial x} \frac{\partial (\delta y)}{\partial x} \right]. \tag{4.37}$$

Quick calculation 4.1 Prove equation (4.37).

We must have no derivatives acting on the variations, so we rewrite each of the two terms above as a full derivative minus a term in which the derivative does not act on the variation:

$$\delta S = \int_{t_i}^{t_f} dt \int_0^a dx \left[\frac{\partial}{\partial t} \left(\mu_0 \frac{\partial y}{\partial t} \delta y \right) - \mu_0 \frac{\partial^2 y}{\partial t^2} \delta y + \frac{\partial}{\partial x} \left(-T_0 \frac{\partial y}{\partial x} \delta y \right) + T_0 \frac{\partial^2 y}{\partial x^2} \delta y \right]. \tag{4.38}$$

The time derivative reduces to evaluations at t_f and t_i, while the space derivative gives evaluations at the string endpoints:

$$\begin{aligned} \delta S = & \int_0^a \left[\mu_0 \frac{\partial y}{\partial t} \delta y \right]_{t=t_i}^{t=t_f} dx + \int_{t_i}^{t_f} \left[-T_0 \frac{\partial y}{\partial x} \delta y \right]_{x=0}^{x=a} dt \\ & - \int_{t_i}^{t_f} dt \int_0^a dx \left(\mu_0 \frac{\partial^2 y}{\partial t^2} - T_0 \frac{\partial^2 y}{\partial x^2} \right) \delta y. \end{aligned} \tag{4.39}$$

Our final expression for δS contains three terms. Each one must vanish independently. The third term, for example, is determined by the motion of the string for $x \in (0, a)$ and $t \in (t_i, t_f)$. In this domain $\delta y(t, x)$ is not restricted by boundary conditions nor by initial or final conditions, so we set to zero the coefficient of δy and recover our original equation (4.6). The first term in (4.39) is determined by the configuration of the string at times t_i and t_f. If we specify these initial and final configurations, we are in effect setting $\delta y(t_i, x)$ and $\delta y(t_f, x)$ to zero. This causes the first term to vanish. We encountered an analogous situation in our study of the free particle.

The second term in (4.39) is new. Written out explicitly, it is

$$\int_{t_i}^{t_f} \left[-T_0 \frac{\partial y}{\partial x}(t,a)\, \delta y(t,a) + T_0 \frac{\partial y}{\partial x}(t,0)\, \delta y(t,0) \right] dt, \tag{4.40}$$

and it concerns the motion of the string endpoints $y(t,0)$ and $y(t,a)$. We need a boundary condition for each of the two terms above. Let x_* denote the x coordinate of an endpoint; x_* can be equal to zero or equal to a. Selecting an endpoint means fixing the value of x_*. We can make each term in (4.40) vanish by specifying either Dirichlet or Neumann boundary conditions. Consider an endpoint x_* and the associated term in (4.40). If we impose a Dirichlet boundary condition the position of the chosen endpoint is fixed throughout time, and we require that the variation $\delta y(t, x_*)$ vanishes. This will cause the chosen term to vanish. If, on the other hand, we assume that the endpoint is free to move, then the variation $\delta y(t, x_*)$ is unconstrained. The term will vanish if we impose the condition

$$\frac{\partial y}{\partial x}(t, x_*) = 0, \quad \text{Neumann boundary condition.} \tag{4.41}$$

Dirichlet boundary conditions can be written in a form where the similarity to Neumann boundary conditions is more apparent. If a string endpoint is fixed, the time derivative of the endpoint coordinate must vanish

$$\frac{\partial y}{\partial t}(t, x_*) = 0, \quad \text{Dirichlet boundary condition.} \tag{4.42}$$

The similarity with (4.41) is quite striking. The only change is that spatial derivatives were turned into time derivatives. If we write Dirichlet boundary conditions in this form, we must still specify the values of the coordinates at the fixed endpoints. In order to appreciate further the physical import of boundary conditions, we consider the momentum p_y carried by the string. There is no other component to the momentum, because we have assumed that the motion is restricted to the y direction. This momentum is simply the sum of the momenta of each infinitesimal segment along the string:

$$p_y = \int_0^a \mu_0 \frac{\partial y}{\partial t}\, dx. \tag{4.43}$$

Let us see if this momentum is conserved:

$$\frac{dp_y(t)}{dt} = \int_0^a \mu_0 \frac{\partial^2 y}{\partial t^2}\, dx = \int_0^a T_0 \frac{\partial^2 y}{\partial x^2}\, dx = T_0 \left[\frac{\partial y}{\partial x} \right]_{x=0}^{x=a}, \tag{4.44}$$

where we used the wave equation (4.6). We see that momentum is conserved when Neumann boundary conditions (4.41) apply at both endpoints. For Dirichlet boundary conditions momentum is not generally conserved! Indeed, when the endpoints of a string are attached to a wall, the wall is constantly exerting a force on the string. In the lowest

normal mode of a Dirichlet string, for example, the net momentum constantly oscillates between the $+y$ and $-y$ directions.

Why is this important for string theory? For a long time string theorists did not take seriously the possibility of Dirichlet boundary conditions. It seemed unphysical that the string momentum could fail to be conserved. Moreover, what could the endpoints of open strings be attached to? The answer is that they are attached to D-branes – a new kind of dynamical extended object. If a string is attached to a D-brane then momentum can be conserved – the momentum lost by the string is absorbed by the D-brane. A detailed analysis of the spatial boundary term induced by variation is crucial to recognize the possibility of D-branes in string theory.

We conclude this chapter with a more general derivation of the equation of motion for the string. For this, we use (4.35) to write the action as

$$S = \int_{t_i}^{t_f} dt \int_0^a dx\, \mathcal{L}\left(\frac{\partial y}{\partial t}, \frac{\partial y}{\partial x}\right). \tag{4.45}$$

We also define the quantities

$$\mathcal{P}^t \equiv \frac{\partial \mathcal{L}}{\partial \dot{y}}, \quad \mathcal{P}^x \equiv \frac{\partial \mathcal{L}}{\partial y'}, \tag{4.46}$$

with $y' = \partial y/\partial x$. These are simply the derivatives of $\mathcal{L}$ with respect to its first and second arguments, respectively. Explicitly, they are

$$\mathcal{P}^t = \mu_0 \frac{\partial y}{\partial t}, \quad \mathcal{P}^x = -T_0 \frac{\partial y}{\partial x}. \tag{4.47}$$

When we vary the motion by δy, the variation of the action is given by

$$\delta S = \int_{t_i}^{t_f} dt \int_0^a dx \left[\frac{\partial \mathcal{L}}{\partial \dot{y}}\, \delta \dot{y} + \frac{\partial \mathcal{L}}{\partial y'}\, \delta y'\right] = \int_{t_i}^{t_f} dt \int_0^a dx \left[\mathcal{P}^t\, \delta \dot{y} + \mathcal{P}^x\, \delta y'\right]. \tag{4.48}$$

Using the standard manipulations we find

$$\begin{aligned} \delta S = &\int_0^a \left[\mathcal{P}^t \delta y\right]_{t=t_i}^{t=t_f} dx + \int_{t_i}^{t_f} \left[\mathcal{P}^x\, \delta y\right]_{x=0}^{x=a} dt \\ &- \int_{t_i}^{t_f} dt \int_0^a dx \left(\frac{\partial \mathcal{P}^t}{\partial t} + \frac{\partial \mathcal{P}^x}{\partial x}\right) \delta y. \end{aligned} \tag{4.49}$$

Quick calculation 4.2 Derive equation (4.49).

Quick calculation 4.3 Match in detail equations (4.49) and (4.39).

The variation in (4.49) gives the equation of motion

$$\frac{\partial \mathcal{P}^t}{\partial t} + \frac{\partial \mathcal{P}^x}{\partial x} = 0. \tag{4.50}$$

Using (4.47) we readily see that this is the wave equation (4.6).

Note that $\mathcal{P}^t$, as given in (4.47), coincides with the momentum density in equation (4.43). This is not an accident. In Lagrangian mechanics, the derivative of the Lagrangian with respect to a velocity is the conjugate momentum. For the string, $\dot{y}$ plays the role of a velocity, so $\mathcal{P}^t$, the derivative of the Lagrangian density with respect to $\dot{y}$, is a momentum density.

In addition, note that for string endpoints that are free to move, the vanishing of δS requires $\mathcal{P}^x = 0$. As we can see from (4.47), this is a Neumann boundary condition. Furthermore, $\mathcal{P}^t$ vanishes at the string endpoints for a Dirichlet boundary condition (4.42). A more detailed analysis of these facts will be given in Chapter 8, where $\mathcal{P}^t$ and $\mathcal{P}^x$ will be shown to have an interesting two-dimensional interpretation.

Problems

Problem 4.1 Consistency of small transverse oscillations.

Reconsider the analysis of transverse oscillations in Section 4.1. Calculate the horizontal force dF_h on the little piece of string shown in Figure 4.1. Show that for small oscillations this force is much smaller than the vertical force dF_v responsible for the transverse oscillations.

Problem 4.2 Longitudinal waves on strings.

Consider a string with uniform mass density μ_0 stretched between $x = 0$ and $x = a$. Let the equilibrium tension be T_0. Longitudinal waves are possible if the tension of the string varies as it stretches or compresses. For a piece of this string with equilibrium length L, a small change ΔL of its length is accompanied by a small change ΔT of the tension where

$$\frac{1}{\tau_0} \equiv \frac{1}{L}\frac{\Delta L}{\Delta T}.$$

Here τ_0 is a tension coefficient with units of tension. Find the equation governing the small longitudinal oscillations of this string. Give the velocity of the waves.

Problem 4.3 A configuration with two joined strings.

A string with tension T_0 is stretched from $x = 0$ to $x = 2a$. The part of the string $x \in (0, a)$ has constant mass density μ_1, and the part of the string $x \in (a, 2a)$ has constant mass density μ_2. Consider the differential equation (4.20) that determines the normal oscillations.

(a) What boundary conditions should be imposed on $y(x)$ and $\frac{dy}{dx}(x)$ at $x = a$?
(b) Write the conditions that determine the possible frequencies of oscillation.
(c) Calculate the lowest frequency of oscillation of this string when $\mu_1 = \mu_0$ and $\mu_2 = 2\mu_0$.

Problem 4.4 Evolving an initial open string configuration.

A string with tension T_0, mass density μ_0, and wave velocity $v_0 = \sqrt{T_0/\mu_0}$, is stretched from $(x, y) = (0, 0)$ to $(x, y) = (a, 0)$. The string endpoints are fixed, and the string can vibrate in the y direction.

(a) Write $y(t, x)$ as in (4.11), and prove that the above Dirichlet boundary conditions imply

$$h_+(u) = -h_-(-u) \quad \text{and} \quad h_+(u) = h_+(u + 2a). \tag{1}$$

Here $u \in (-\infty, \infty)$ is a dummy variable that stands for the argument of the functions $h_\pm$.

Now consider an initial value problem for this string. At $t = 0$ the transverse displacement is identically zero, and the velocity is

$$\frac{\partial y}{\partial t}(0, x) = v_0 \, \frac{x}{a}\Big(1 - \frac{x}{a}\Big), \quad x \in (0, a). \tag{2}$$

(b) Calculate $h_+(u)$ for $u \in (-a, a)$. Does this define $h_+(u)$ for all u?

(c) Calculate $y(t, x)$ for x and $v_0 t$ in the domain D defined by the two conditions

$$D = \{(x, v_0 t) \big| 0 \le x \pm v_0 t < a\}.$$

Exhibit the domain D in a plane with axes x and $v_0 t$.

(d) At $t = 0$ the midpoint $x = a/2$ has the largest velocity of all points in the string. Show that the velocity of the midpoint reaches the value of zero at time $t_0 = a/(2v_0)$ and that $y(t_0, a/2) = a/12$. This is the maximum vertical displacement of the string.

Problem 4.5 Closed string motion.

We can describe a nonrelativistic closed string fairly accurately by having the string wrapped around a cylinder of large circumference $2\pi R$ on which it is kept taut by the string tension T_0. We assume that the string can move on the surface of the cylinder without experiencing any friction. Let x be a coordinate along the circumference of the cylinder: $x \sim x + 2\pi R$ and let y be a coordinate perpendicular to x, thus running parallel to the axis of the cylinder. As expected, the general solution for transverse motion is given by

$$y(x, t) = h_+(x - v_0 t) + h_-(x + v_0 t),$$

where $h_+(u)$ and $h_-(v)$ are arbitrary functions of single variables u and v with $-\infty < u, v < \infty$. The string has mass per unit length μ_0, and $v_0 = \sqrt{T_0/\mu_0}$.

(a) State the periodicity condition that must be satisfied by $y(x, t)$ on account of the identification that applies to the x coordinate. Show that the *derivatives* $h'_+(u)$ and $h'_-(v)$ are, respectively, periodic functions of u and v.

(b) Show that one can write

$$h_+(u) = \alpha u + f(u), \quad h_-(v) = \beta v + g(v),$$

where f and g are periodic functions and α and β are constants. Give the relation between α and β that follows from (a).

(c) Calculate the total momentum carried by the string in the y direction. Is it conserved?

Problem 4.6 Stationary action: minima and saddles.

A particle perfoming harmonic motion along the x axis can be used to show that classical solutions are not always minima of the action functional. The action for this particle is

$$S[x] = \int_0^{t_f} L\, dt = \int_0^{t_f} dt\, \frac{1}{2} m^2 \left(\dot{x}^2 - \bar{\omega}^2 x^2\right),$$

where m is the mass of the particle, $\bar{\omega}$ is the frequency of oscillation, and the motion happens for $t \in [0, t_f]$. Consider a classical solution $\bar{x}(t)$ and a variation $\delta x(t)$ that vanishes at $t = 0$ and $t = t_f$.

(a) Show that the variation of the action is *exactly* given by

$$\Delta S[\delta x] \equiv S[\bar{x} + \delta x] - S[\bar{x}] = \frac{1}{2} m^2 \int_0^{t_f} dt \left(\left(\frac{d\delta x}{dt} \right)^2 - \bar{\omega}^2 \delta x^2 \right).$$

It is noteworthy that ΔS only depends on δx; $\bar{x}$ drops out from the answer.

(b) A complete set of variations that vanish at $t = 0$ and $t = t_f$ takes the form

$$\delta_n x = \sin \omega_n t\,, \quad \text{with} \quad \omega_n = \frac{\pi n}{t_f} \quad \text{and} \quad n = 1, 2, \ldots, \infty\,.$$

The general variation δx that vanishes at $t = 0$ and $t = t_f$ is a linear superposition of variations $\delta_n x$ with arbitrary coefficients b_n. Calculate $\Delta S[\delta_n x]$ (your answer should vanish for $\omega_n = \bar{\omega}$). Prove that

$$\Delta S\Big[\sum_{n=1}^{\infty} b_n \delta_n x\Big] = \sum_{n=1}^{\infty} \Delta S\big[b_n \delta_n x\big]\,.$$

(c) Show that for $t_f < \frac{\pi}{\bar{\omega}}$ one gets $\Delta S[\delta_n x] > 0$ for all $n \geq 1$. Explain why this guarantees that the classical solution is a minimum of the action. Show that for $\frac{\pi}{\bar{\omega}} < t_f < \frac{2\pi}{\bar{\omega}}$ all variations $\delta_n x$ lead to $\Delta S > 0$, except for $\delta_1 x$, which leads to $\Delta S < 0$. In this case the classical solution is a saddle point: there are variations that increase the action and variations that decrease the action. As t_f increases, the number of variations $\delta_n x$ that decrease the action increases.

Problem 4.7 Variational problem for strings.

Consider a string stretched from $x = 0$ to $x = a$, with a tension T_0 and a position-dependent mass density $\mu(x)$. The string is fixed at the endpoints and can vibrate in the y direction. Equation (4.20) determines the oscillation frequencies ω_i and associated profiles $\psi_i(x)$ for this string.

(a) Set up a variational procedure that gives an upper bound on the lowest frequency of oscillation ω_0. (This can be done roughly as in quantum mechanics, where the ground

state energy E_0 of a system with Hamiltonian H satisfies $E_0 \leq (\psi, H\psi)/(\psi, \psi)$.) As a useful first step consider the inner product

$$(\psi_i, \psi_j) \equiv \int_0^a \mu(x)\psi_i(x)\psi_j(x)dx$$

and show that it vanishes when $\omega_i \neq \omega_j$. Explain why your variational procedure works.

(b) Consider the case $\mu(x) = \mu_0 \frac{x}{a}$. Use your variational principle to find a simple bound on the lowest oscillation frequency. Compare with the answer $\omega_0^2 \simeq (18.956)\frac{T_0}{\mu_0 a^2}$ obtained by a direct numerical solution of the eigenvalue problem.

Problem 4.8 Deriving Euler–Lagrange equations.[†]

(a) Consider an action for a dynamical variable $q(t)$:

$$S = \int dt\, L(q(t), \dot{q}(t); t). \tag{1}$$

Calculate the variation δS of the action under a variation $\delta q(t)$ of the coordinate. Use the condition $\delta S = 0$ to find the equation of motion for the coordinate $q(t)$ (the Euler–Lagrange equation).

(b) Consider an action for a dynamical field variable $\phi(t, \vec{x})$. As indicated, the field is a function of space and time, and is briefly written as the spacetime function $\phi(x)$. The action is obtained by integrating the Lagrangian density $\mathcal{L}$ over spacetime. The Lagrangian density is a function of the field and the spacetime derivatives of the field:

$$S = \int d^D x\ \mathcal{L}(\phi(x), \partial_\mu \phi(x)). \tag{2}$$

Here $d^D x = dt\, dx^1 \ldots dx^d$, and $\partial_\mu \phi = \partial\phi/\partial x^\mu$. Calculate the variation δS of the action under a variation $\delta\phi(x)$ of the field. Use the condition $\delta S = 0$ to find the equation of motion for the field $\phi(x)$ (the Euler–Lagrange equation).

5 The relativistic point particle

To formulate the dynamics of a system we can write either the equations of motion or, alternatively, an action. In the case of the relativistic point particle it is rather easy to write the equations of motion. But the action is so physical and geometrical that it is worth pursuing in its own right. More importantly, while it is difficult to guess the equations of motion for the relativistic string, the action is a natural generalization of the relativistic particle action that we will study in this chapter. We conclude with a discussion of the charged relativistic particle.

5.1 Action for a relativistic point particle

In this section we learn how to formulate the relativistic theory that describes a free point particle of rest mass $m > 0$. A free particle is a particle that is not subject to any force. Our analysis begins with some preliminary remarks about units and nonrelativistic particles.

For any dynamical system, the action S is obtained by integrating the Lagrangian over time. Since the Lagrangian has units of energy, the action has units of energy times time:

$$[S] = M\,\frac{L^2}{T^2}\,T = \frac{ML^2}{T}. \tag{5.1}$$

It is worth noting that the action has the same units as $\hbar$. Indeed, a form of the quantum mechanical uncertainty principle states that the product of energy and time uncertainties is of order $\hbar$.

The action $S_{\rm nr}$ for a free *nonrelativistic* particle is given by the time integral of the kinetic energy:

$$S_{\rm nr} = \int L_{\rm nr}\,dt = \int \frac{1}{2}mv^2(t)\,dt, \quad v^2 \equiv \vec{v}\cdot\vec{v}, \quad \vec{v} = \frac{d\vec{x}}{dt}, \quad v = |\vec{v}|. \tag{5.2}$$

The equation of motion which follows by Hamilton's principle is

$$\frac{d\vec{v}}{dt} = 0. \tag{5.3}$$

The free particle moves with constant velocity. Since even a free *relativistic* particle must move with constant velocity, how do we know that the action $S_{\rm nr}$ is not correct in relativity? Perhaps the simplest answer is that this action allows the particle to move with *any* constant

velocity, even one that exceeds the velocity of light. The velocity of light does not even appear in this action. $S_{\rm nr}$ cannot be the action for a relativistic point particle.

We now construct a relativistic action S for the free point particle. We will do this by making an educated guess and then showing that it works properly. Since we are interested in relativistic physics it is convenient to represent the motion of the particle in *spacetime*. The path traced out by the particle in spacetime is called the *world-line* of the particle. Even a static particle traces a line in spacetime since time always flows.

A physically consistent action must yield Lorentz invariant equations of motion. Let us elaborate on this point. Suppose that a particular Lorentz observer tells you that a particle appears to be moving in accordance to its equations of motion, plainly, that the particle is performing physical motion. Then, you should expect that any other Lorentz observer will tell you that the particle is doing physical motion. It would be inconsistent for one observer to state that a certain motion is allowed and for another observer to state that the same motion is forbidden. If the equations of motion hold in a fixed Lorentz frame, they must hold in all Lorentz frames. This is what Lorentz invariance of the equations of motion means.

We are going to write an action, and we are going to take our time to find the equations of motion. Is there any way to impose a constraint on the action that will result in the Lorentz invariance of the equations of motion? Yes, there is. We require the action to be a Lorentz scalar: for *any* particle world-line, all Lorentz observers must compute the same value for the action. Since the action has no spacetime indices, this is a reasonable requirement. If the action is a Lorentz scalar, the equations of motion will be Lorentz invariant. The reason is simple and neat. Suppose one Lorentz observer states that, for a given world-line, the action is stationary against all variations of the world-line. Since all Lorentz observers agree on the value of the action for any world-line, they will all agree that the action is stationary about the world-line in question. By Hamilton's principle, the world-line that makes the action stationary satisfies the equations of motion, and therefore all Lorentz observers will agree that the equations of motion are satisfied for the world-line in question.

Lorentz invariance imposes strong constraints on the possible forms of the action. In fact, there are valid grounds to worry that Lorentz invariance is too strong a constraint on the action. The nonrelativistic action in (5.2), for example, is *not* invariant under a Galilean boost $\vec{v} \to \vec{v} + \vec{v}_0$ with constant $\vec{v}_0$. Such a boost is a symmetry of the theory, since the equation of motion (5.3) *is* invariant. Similarly, it could happen that the equations of motion for the relativistic point particle are Lorentz invariant but that the action is not. Fortunately, this complication does not occur in this case; we will find a satisfactory fully Lorentz invariant action.

Quick calculation 5.1 Calculate explicitly the variation of the action $S_{\rm nr}$ under a boost.

We know that the action is a functional – it takes as input a set of functions that describe a world-line and it outputs a number S. Imagine a particle whose spacetime trajectory starts at the origin and ends at $(ct_f, \vec{x}_f)$. There are many possible world-lines between the starting and ending points, as shown in Figure 5.1 (which uses one spatial dimension

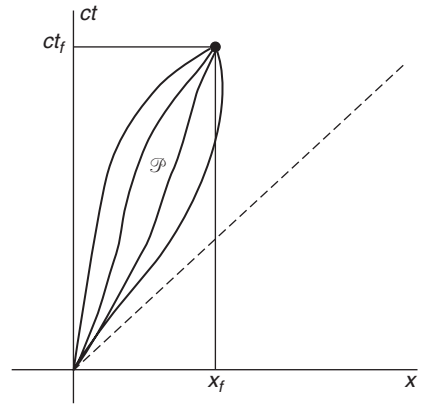

Fig. 5.1 **A spacetime diagram with a series of world-lines connecting the origin to the spacetime point (ct_f, x_f).**

for ease of representation). We would like that, for any world-line, all Lorentz observers compute the same value for the action. Let $\mathcal{P}$ denote one world-line. What quantity related to $\mathcal{P}$ do all Lorentz observers agree on? The elapsed proper time! All Lorentz observers agree on the amount of time that elapses on a clock carried by the moving particle. So let us take the action of the world-line $\mathcal{P}$ to be proportional to the proper time associated with it.

To formulate this idea quantitatively, we recall that

$$-ds^2 = -c^2dt^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2, \tag{5.4}$$

and that the infinitesimal proper time is equal to ds/c (recall that $ds^2 = (ds)^2$ since intervals are timelike). The integral of (ds/c) over the path $\mathcal{P}$ gives the proper time elapsed on $\mathcal{P}$. Since proper time has units of time, to get the units of action we need an additional multiplicative factor with units of energy or units of mass times velocity-squared. This factor should be Lorentz invariant, to preserve the Lorentz invariance of our partial guess (ds/c). For the mass we can use m, the rest mass of the particle, and for the velocity we can use c, the fundamental velocity in relativity. We cannot use the particle velocity because it is not a Lorentz invariant. The factor is then mc^2, which is, in fact, the rest energy of the particle. Therefore, our guess for the action is the integral of $mc^2\,(ds/c) = mc\,ds$. Of course, there is still the possibility that a dimensionless numerical factor is missing. It turns out that there should be a minus sign, but the unit coefficient is correct. We therefore claim that the correct action is

$$S = -mc \int_{\mathcal{P}} ds. \tag{5.5}$$

The action is equal to minus the rest energy times the proper time. This action is so simple looking that it may be baffling. It probably looks nothing like the actions you have seen before. We can make its content more familiar by choosing a particular Lorentz observer

and expressing the action as the integral of a Lagrangian over time. With the help of (5.4) we relate ds to dt by

$$ds = c\,dt\sqrt{1-\frac{v^2}{c^2}}. \tag{5.6}$$

This allows us to write the action in (5.5) as an integral over time:

$$S = -mc^2\int_{t_i}^{t_f} dt\sqrt{1-\frac{v^2}{c^2}}, \tag{5.7}$$

where t_i and t_f are the values of time at the initial and final points of the world-line $\mathcal{P}$, respectively. From this version of the action, we see that the relativistic Lagrangian for the point particle is

$$L = -mc^2\sqrt{1-\frac{v^2}{c^2}}. \tag{5.8}$$

The Lagrangian is equal to minus the rest energy times a relativistic factor. This Lagrangian makes no sense when $v > c$ since it ceases to be real. The constraint of maximal velocity is therefore implemented. This could have been anticipated: proper time is only defined for motion where the velocity does not exceed the velocity of light. The paths shown in Figure 5.1 all represent motion where the velocity of the particle never exceeds the velocity of light. Only for such paths is the action defined. At any point in any of those paths, the tangent vector to the path is a timelike vector.

To show that this Lagrangian gives the familiar physics in the limit of small velocities, we expand the square root assuming $v \ll c$. Keeping just the first term in the expansion gives

$$L \simeq -mc^2\Big(1-\frac{1}{2}\frac{v^2}{c^2}\Big) = -mc^2+\frac{1}{2}mv^2. \tag{5.9}$$

Constant terms in a Lagrangian do not affect the equations of motion, so the term $(-mc^2)$ can be ignored for this purpose. The leading significant term coincides with the nonrelativistic Lagrangian in (5.2), showing that the familiar nonrelativistic physics emerges. This also confirms that we normalized the relativistic Lagrangian correctly.

The canonical momentum is the derivative of the Lagrangian with respect to the velocity. Using (5.8) we find

$$\vec{p} = \frac{\partial L}{\partial \vec{v}} = -mc^2\Big(-\frac{\vec{v}}{c^2}\Big)\frac{1}{\sqrt{1-v^2/c^2}} = \frac{m\vec{v}}{\sqrt{1-v^2/c^2}}. \tag{5.10}$$

This is just the relativistic momentum of the point particle. What about the Hamiltonian? It is given by

$$H = \vec{p}\cdot\vec{v} - L = \frac{mv^2}{\sqrt{1-v^2/c^2}} + mc^2\sqrt{1-v^2/c^2} = \frac{mc^2}{\sqrt{1-v^2/c^2}}, \tag{5.11}$$

where the result was left as a function of the velocity of the particle, rather than as a function of its momentum. As expected, the answer coincides with the relativistic energy (2.68) of the point particle.

We have therefore recovered the familiar physics of a relativistic particle from the rather remarkable action (5.5). This action is very elegant: it is briefly written in terms of the geometrical quantity ds, it has a clear physical interpretation as total proper time, and it manifestly guarantees the Lorentz invariance of the physics it describes.

5.2 Reparameterization invariance

In this section we explore an important property of the point particle action (5.5). This property is called reparameterization invariance. To evaluate the integral in the action, an observer may find it useful to parameterize the particle world-line. Reparameterization invariance of the action means that the value of the action is independent of the parameterization chosen to calculate it. This should be so, since the action (5.5) is in fact defined independently of any parameterization: the integration can be done by breaking $\mathcal{P}$ into small pieces and adding the values of $mc\,ds$ for each piece. No parameterization is needed to do this. In practice, however, world-lines are described as parameterized lines, and the parameterization *is* used to compute the action.

We parameterize the world-line $\mathcal{P}$ of a point particle using a parameter τ (Figure 5.2). This parameter must be strictly increasing as the world-line goes from the initial point x_i^μ to the final point x_f^μ, but is otherwise arbitrary. As τ ranges in the interval $[\tau_i, \tau_f]$ it describes the motion of the particle. To have a parameterization of the world-line means that we have expressions for the coordinates x^μ as functions of τ:

$$x^\mu = x^\mu(\tau). \tag{5.12}$$

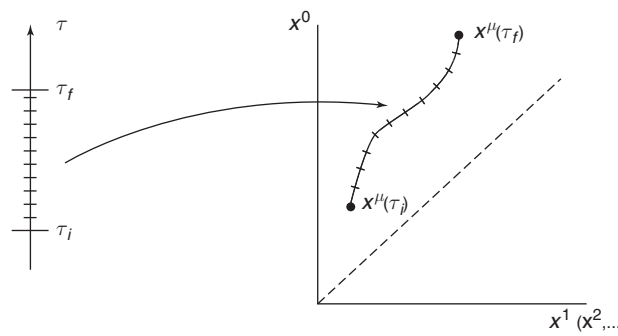

Fig. 5.2 A world-line fully parameterized by τ. All spacetime coordinates x^μ are functions of τ.

We also require

$$x_i^\mu = x^\mu(\tau_i), \quad x_f^\mu = x^\mu(\tau_f). \tag{5.13}$$

Note that even the time coordinate x^0 is parameterized. Normally, we use time as a parameter and describe position as a function of time. This is what we did in Section 5.1. But if we want to treat space and time coordinates on the same footing, we must parameterize both in terms of an additional parameter τ.

We now reexpress the integrand ds using the parameterized world-line. To this end, we use $ds^2 = -\eta_{\mu\nu}dx^\mu dx^\nu$ to write

$$ds^2 = -\eta_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau}(d\tau)^2. \tag{5.14}$$

For any motion where the velocity does not exceed the velocity of light $ds^2 = (ds)^2$, and therefore the action (5.5) takes the form

$$S = -mc\int_{\tau_i}^{\tau_f}\sqrt{-\eta_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau}}\;d\tau. \tag{5.15}$$

This is the explicit form of the action when the path has been parameterized by τ.

We have already seen that the value of the action is the same for all Lorentz observers. We have now fixed an observer, who has calculated the action using some parameter τ. Does the value of the action depend on the choice of parameter? It does not. The observer can reparameterize the world-line, and the value of the action will be the same. Thus S is *reparameterization invariant*. To see this, suppose we change the parameter from τ to τ'. Then, by the chain rule,

$$\frac{dx^\mu}{d\tau} = \frac{dx^\mu}{d\tau'}\frac{d\tau'}{d\tau}. \tag{5.16}$$

Substituting back into (5.15), we get

$$S = -mc\int_{\tau_i}^{\tau_f}\sqrt{-\eta_{\mu\nu}\frac{dx^\mu}{d\tau'}\frac{dx^\nu}{d\tau'}}\;\frac{d\tau'}{d\tau}d\tau = -mc\int_{\tau_i'}^{\tau_f'}\sqrt{-\eta_{\mu\nu}\frac{dx^\mu}{d\tau'}\frac{dx^\nu}{d\tau'}}\,d\tau', \tag{5.17}$$

which has the same form as (5.15), thus establishing the reparameterization invariance. Because the verification of this property is quite simple, we say that the action (5.15) is *manifestly* reparameterization invariant.

5.3 Equations of motion

We now move on to the equations of motion. For this we must calculate the variation δS of the action (5.5) when the world-line of the particle is varied by a small amount $\delta x^\mu(\tau)$. Here τ is an arbitrary parameter along the path. The variation is simply given by

$$\delta S = -mc\int \delta(ds). \tag{5.18}$$

The variation of ds can be found from the simpler variation of $ds^2 = (ds)^2$. Varying both sides of (5.14) we find

$$2\,ds\,\delta(ds) = -2\eta_{\mu\nu}\,\delta\Big(\frac{dx^\mu}{d\tau}\Big)\,\frac{dx^\nu}{d\tau}(d\tau)^2. \tag{5.19}$$

The factor of two on the right-hand side arises because, by symmetry, the variations of $\frac{dx^\mu}{d\tau}$ and $\frac{dx^\nu}{d\tau}$ give the same result. Since the variation of a velocity is equal to the time derivative of the variation of the coordinate,

$$\delta\Big(\frac{dx^\mu}{d\tau}\Big) = \frac{d(\delta x^\mu)}{d\tau}. \tag{5.20}$$

Using this result and simplifying (5.19) a bit, we get

$$\delta(ds) = -\eta_{\mu\nu}\frac{d(\delta x^\mu)}{d\tau}\frac{dx^\nu}{ds}\,d\tau = -\frac{d(\delta x^\mu)}{d\tau}\frac{dx_\mu}{ds}\,d\tau, \tag{5.21}$$

where $\eta_{\mu\nu}$ was used to lower the index of dx^ν. We can now go ahead and vary the action using (5.18):

$$\delta S = mc\int_{\tau_i}^{\tau_f}\frac{d(\delta x^\mu)}{d\tau}\,\frac{dx_\mu}{ds}\,d\tau. \tag{5.22}$$

Here we introduced explicit limits to the integration: τ_i and τ_f denote the values of the parameter at the initial and final points of the world-line, respectively. We recognize that

$$mc\frac{dx_\mu}{ds} = mu_\mu = p_\mu, \tag{5.23}$$

and, as a result, the variation of the action takes the form

$$\delta S = \int_{\tau_i}^{\tau_f}\frac{d(\delta x^\mu)}{d\tau}\;p_\mu\;d\tau. \tag{5.24}$$

To get an equation of motion we need to have δx^μ multiplying an object under the integral – the equation of motion is then simply the vanishing of that object. Since there are still derivatives acting on δx^μ, we rewrite the integrand as a total derivative plus additional terms where δx^μ appears multiplicatively:

$$\delta S = \int_{\tau_i}^{\tau_f} d\tau\,\frac{d}{d\tau}\Big(\delta x^\mu\,p_\mu\Big) - \int_{\tau_i}^{\tau_f} d\tau\;\delta x^\mu(\tau)\;\frac{dp_\mu}{d\tau}. \tag{5.25}$$

The first integral gives $\delta x^\mu\,p_\mu$ evaluated at the boundaries of the world-line. This term vanishes because we fix the coordinates on the boundaries. Since the second term must vanish for arbitrary $\delta x^\mu(\tau)$, we obtain the equation of motion

$$\frac{dp_\mu}{d\tau} = 0. \tag{5.26}$$

It is clear that $dp^\mu/d\tau$ also vanishes. The equation of motion states that the momentum p_μ (or p^μ) of the point particle is constant along its world-line. This is a parameterization-independent statement. It implies, of course, that the momentum is constant in time. If a

function is constant over a line, its derivative with respect to any parameter used to describe the line will vanish. Indeed, the parameter τ in (5.26) is arbitrary. We obtained this equation by varying the relativistic action for the point particle using fully relativistic notation.

Quick calculation 5.2 Show that equation (5.26) implies that

$$\frac{dp_\mu}{d\tau'} = 0 \tag{5.27}$$

holds for an arbitrary parameter $\tau'(\tau)$. What should be true about $d\tau'/d\tau$ for τ' to be a good parameter when τ is one?

If we parameterize the world-line with the proper time s, equation (5.26) gives

$$\frac{dp^\mu}{ds} = 0. \tag{5.28}$$

Using (5.23) to write the momentum as a derivative of the position with respect to proper time, we find

$$\frac{d^2x^\mu}{ds^2} = 0. \tag{5.29}$$

This is an equivalent version of the equation of motion. The constancy of dx^μ/ds means that on a path marked by equal intervals of proper time, the change in x^μ between any successive pair of marks is the same. Equation (5.29) does *not* hold when s is replaced by an arbitrary parameter τ. This is reasonable: an arbitrary parameter means arbitrarily spaced marks, so the change in x^μ between any successive pair of new marks need not be the same. It is actually possible to write a slightly more complicated version of (5.29) that uses an arbitrary parameter and is manifestly reparameterization invariant (Problem 5.2).

Our goal in this section has been achieved: we have shown how to derive the physically expected equation of motion (5.29) (or (5.26)), starting from the Lorentz invariant action (5.5). As we explained earlier, the resulting equation of motion is guaranteed to be Lorentz invariant. Let us check this explicitly.

Under a Lorentz transformation, the coordinates x^μ transform as indicated in equation (2.38): $x'^\mu = L^\mu{}_\nu\, x^\nu$, where the constants $L^\mu{}_\nu$ can be viewed as the entries of an invertible matrix L. Since ds is the same in all Lorentz frames, the equation of motion in primed coordinates is (5.29), with x^μ replaced by x'^μ:

$$0 = \frac{d^2x'^\mu}{ds^2} = \frac{d^2}{ds^2}(L^\mu{}_\nu\, x^\nu) = L^\mu{}_\nu\, \frac{d^2x^\nu}{ds^2}. \tag{5.30}$$

Since the matrix L is invertible, the above equation implies equation (5.29). Namely, if the equation of motion holds in the primed coordinates, it holds in the unprimed coordinates as well. This is the Lorentz invariance of the equations of motion.

5.4 Relativistic particle with electric charge

The point particle we have considered so far is free and it moves with constant four-velocity or four-momentum. If a point particle is electrically charged and there are nontrivial electromagnetic fields, the particle will experience forces and its four-momentum will not be constant. You know, in fact, how the momentum of such a particle varies in time. Its time derivative is governed by the Lorentz force equation (3.5), which was written in relativistic notation in Problem 3.1:

$$\frac{dp_\mu}{ds} = \frac{q}{c} F_{\mu\nu} \frac{dx^\nu}{ds}. \tag{5.31}$$

This is a relatively intricate equation which involves the field strength and the four-velocity of the particle. Since ds appears on both sides of the equation, the equation in fact holds for a general parameter τ:

$$\frac{dp_\mu}{d\tau} = \frac{q}{c} F_{\mu\nu} \frac{dx^\nu}{d\tau}. \tag{5.32}$$

In the spirit of our previous analysis we try to write an action that gives this equation upon variation. The action turns out to be remarkably simple.

Since the Maxwell field couples to the point particle along its world-line $\mathcal{P}$, we should add to the action (5.5) an integral over $\mathcal{P}$ representing the interaction of the particle with the electromagnetic field. The integral must be Lorentz invariant, and the form of (5.32) suggests that it involves the four-velocity of the particle. Since the four-velocity has one spacetime index, to obtain a Lorentz scalar we must multiply it against another object with one index. The natural candidate is the gauge potential A_μ. We claim that the interaction term in the action is

$$\frac{q}{c} \int_{\mathcal{P}} d\tau \; A_\mu(x(\tau)) \frac{dx^\mu}{d\tau}(\tau). \tag{5.33}$$

Here q is the electric charge, and the integral is over the world-line $\mathcal{P}$, parameterized with the arbitrary parameter τ. At each τ, the vector $(dx^\mu/d\tau)$ is dot multiplied against the gauge potential A_μ, evaluated at the position $x(\tau)$ of the particle. The integrand can be written more briefly as $A_\mu dx^\mu$, by cancelling the factors of $d\tau$. In this form, the interaction term is manifestly independent of parameterization. The world-line of the particle is a one-dimensional space, and the natural field that can couple to a particle in a Lorentz invariant way is a field with one index. This will have an interesting generalization when we consider the motion of strings. Since strings are one-dimensional, they trace out two-dimensional world-sheets in spacetime. We will see that they couple naturally to fields with two Lorentz indices!

The full action for the electrically charged point particle is obtained by adding the term in (5.33) to (5.5):

$$S = -mc \int_{\mathcal{P}} ds + \frac{q}{c} \int_{\mathcal{P}} A_\mu(x) dx^\mu. \tag{5.34}$$

This Lorentz invariant action is simple and elegant. The equation of motion (5.32) arises by setting to zero the variation of S under a change δx^μ of the particle world-line. I do not want to take away from you the satisfaction of deriving this important result. I have therefore left to Problem 5.5 the task of varying the action (5.34) and deriving the equation of motion.

Problems

Problem 5.1 Point particle equation of motion and reparameterizations.

If the path of a point particle is parameterized by proper time, the equation of motion is (5.29). Consider now a new parameter $\tau = f(s)$. Find the most general function f for which (5.29) implies

$$\frac{d^2 x^\mu}{d\tau^2} = 0.$$

Problem 5.2 Particle equation of motion with arbitrary parameterization.

Vary the point particle action (5.15) to find a *manifestly* reparameterization invariant form of the free particle equation of motion.

Problem 5.3 Current of a charged point particle.

Consider a point particle with charge q whose motion in a $D = d + 1$-dimensional space-time is described by functions $x^\mu(\tau) = \{x^0(\tau), \vec{x}(\tau)\}$, where τ is a parameter. The moving particle generates an electromagnetic current $j^\mu = (c\rho, \vec{j})$.

(a) Use delta functions to write expressions for the current components $j^0(\vec{x}, t)$ and $j^i(\vec{x}, t)$.

(b) Show that your answers in (a) arise from the integral representation

$$j^\mu(t, \vec{x}) = qc \int d\tau \, \delta^D(x - x(\tau)) \, \frac{dx^\mu(\tau)}{d\tau}.$$

Here $\delta^D(x) \equiv \delta(x^0)\delta(x^1) \ldots \delta(x^d)$.

Problem 5.4 Hamiltonian for a nonrelativistic charged particle.[†]

The action for a nonrelativistic particle of mass m and charge q coupled to an electromagnetic field is obtained by replacing the first term in (5.34) by the nonrelativistic action for a free point particle:

$$S = \int \frac{1}{2} m v^2 \, dt + \frac{q}{c} \int A_\mu(x) \frac{dx^\mu}{dt} dt.$$

We have also chosen to use time to parameterize the second integral.

(a) Rewrite the action S in terms of the potentials $(\Phi, \vec{A})$ and the ordinary velocity $\vec{v}$. What is the Lagrangian?
(b) Calculate the canonical momentum $\vec{p}$ conjugate to the position of the particle and show that it is given by $\vec{p} = m\vec{v} + \frac{q}{c}\vec{A}$.
(c) Show that the Hamiltonian for the charged particle is

$$H = \frac{1}{2m}\left(\vec{p} - \frac{q}{c}\vec{A}\right)^2 + q\,\Phi.$$

Problem 5.5 Equations of motion for a charged point particle.

Consider the variation of the action (5.34) under a variation $\delta x^\mu(x)$ of the particle trajectory. The variation of the first term in the action was obtained in Section 5.3. Vary the second term (written more explicitly in (5.33)) and show that the equation of motion is (5.32). Begin your calculation by explaining why

$$\delta A_\mu(x(\tau)) = \frac{\partial A_\mu}{\partial x^\nu}(x(\tau))\,\delta x^\nu(\tau).$$

Problem 5.6 Electromagnetic field dynamics with a charged particle.†

The action for the dynamics of *both* a charged point particle and the electromagnetic field is given by

$$S' = -mc\int_{\mathcal{P}} ds + \frac{q}{c}\int_{\mathcal{P}} A_\mu(x)dx^\mu - \frac{1}{4c}\int d^D x\ F_{\mu\nu}F^{\mu\nu}.$$

Here $d^D x = dx^0 dx^1 \ldots dx^d$. Note that the action S' is a hybrid; the last term is an integral over spacetime, and the first two terms are integrals over the particle world-line. While included for completeness, the first term will play no role here. Obtain the equation of motion for the electromagnetic field in the presence of the charged particle by calculating the variation of S' under a variation δA_μ of the gauge potential. The answer should be equation (3.34), where the current is the one calculated in Problem 5.3. [Hint: to vary $A_\mu(x)$ in the world-line action it is useful to rewrite this term as a full spacetime integral with the help of delta functions.]

Problem 5.7 Point particle action in curved space.

In Section 3.6 we considered the invariant interval $ds^2 = -g_{\mu\nu}(x)dx^\mu dx^\nu$ in a curved space with metric $g_{\mu\nu}(x)$. The motion of a point particle of mass m on curved space is studied using the action

$$S = -mc\int ds.$$

Show that the equation of motion obtained by variation of the world-line is

$$\frac{d}{ds}\left[g_{\mu\rho}\frac{dx^\mu}{ds}\right] = \frac{1}{2}\frac{\partial g_{\mu\nu}}{\partial x^\rho}\frac{dx^\mu}{ds}\frac{dx^\nu}{ds}.$$

This is called the geodesic equation. When the metric is constant we recover the equation of motion of a free point particle.

6 Relativistic strings

We now begin our study of the classical relativistic string – a string that is, in many ways, much more elegant than the nonrelativistic one considered before. Inspired by the point particle case, we focus our attention on the surface traced out by the string in spacetime. We use the proper area of this surface as the action; this is the Nambu–Goto action. We study the reparameterization property of this action, identify the string tension, and find the equations of motion. For open strings, we focus on the motion of the endpoints and introduce the concept of D-branes. Finally, we see that the only physical motion is transverse to the string.

6.1 Area functional for spatial surfaces

The action for a relativistic string must be a functional of the string trajectory. Just as a particle traces out a line in spacetime, a string traces out a surface. The line traced out by the particle in spacetime is called the world-line. The two-dimensional surface traced out by a string in spacetime will be called the *world-sheet*. A closed string, for example, will trace out a tube, while an open string will trace out a strip. These two-dimensional world-sheets are shown in the spacetime diagram of Figure 6.1. The lines of constant x^0 in these surfaces are the strings. These are the objects an observer sees at the fixed time x^0. They are open curves for the surface describing the open string evolution (left), and they are closed curves for the surface describing the closed string evolution (right).

In Chapter 5 we learned that the point particle action is proportional to the proper time elapsed on the point particle world-line. The proper time, multiplied by c, is the Lorentz invariant "proper length" of the world-line. For strings we will define the Lorentz invariant "proper area" of a world-sheet. The relativistic string action will be proportional to this proper area, and is called the Nambu–Goto action.

Area functionals are useful in other applications: a soap film held between two rings, for example, automatically constructs the surface of minimal area which joins one ring to the other (Figure 6.2). The string world-sheet and the soap bubble between two rings are very different types of surfaces. At any given instant of time a Lorentz observer will see the full two-dimensional surface of the soap film, but he or she can only see one string from the two-dimensional world-sheet. Imagine that the soap film is static in some Lorentz frame. In this case, time is not relevant to the description of the film, and we think of the film as a *spatial surface*, namely, a surface that extends along two spatial dimensions. The surface

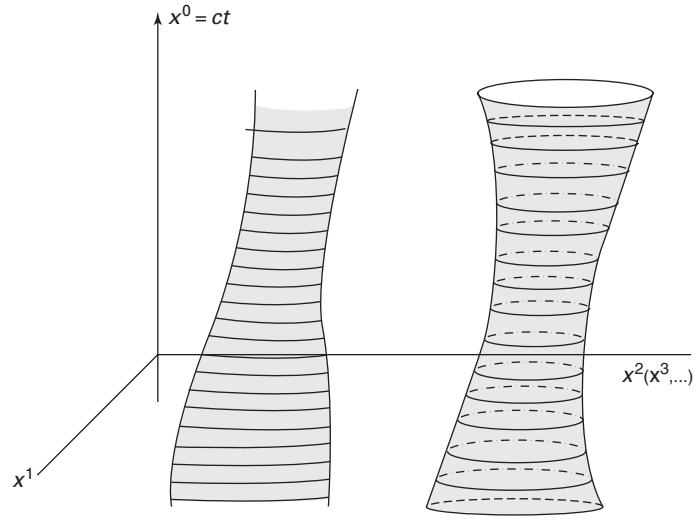

Fig. 6.1 The world-sheets traced out by an open string (left) and by a closed string (right).

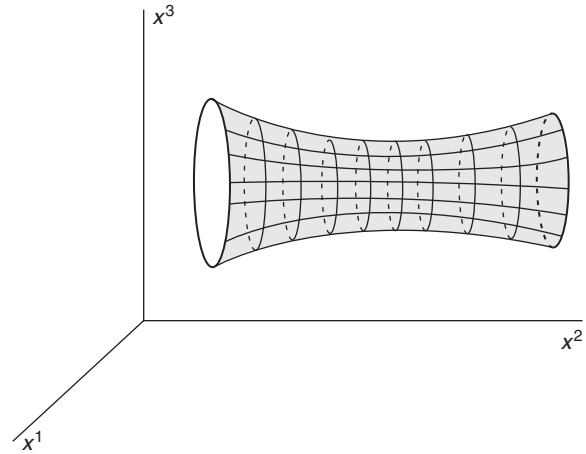

Fig. 6.2 A spatial surface stretching between two rings. If the surface were a soap film, it would be a minimal area surface.

exists in its entirety at any instant of time. We will first study these familiar surfaces, and then we will apply our experience to the case of surfaces in spacetime.

A line in space can be parameterized using only one parameter. A surface in space is two-dimensional, so it requires two parameters ξ^1 and ξ^2. Given a parameterized surface, we can draw on that surface the lines of constant ξ^1 and the lines of constant ξ^2. These lines cover the surface with a grid. We call *target space* the world where the two-dimensional surface lives. In the case of a soap bubble in three dimensions, the target space is the three-dimensional space x^1, x^2, and x^3. The parameterized surface is described by the collection of functions

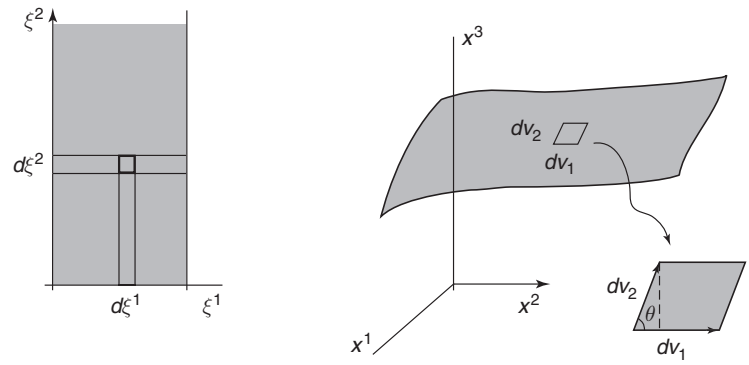

Fig. 6.3 **Left: the parameter space, with a little rectangle selected. Right: the target space surface with the image of the little rectangle, a parallelogram whose sides are the vectors $d\vec{v}_1$ and $d\vec{v}_2$ (shown magnified at the end of the wiggly arrow).**

$$\vec{x}\,(\xi^1, \xi^2) = \left(x^1(\xi^1, \xi^2)\,,\ x^2(\xi^1, \xi^2)\,,\ x^3(\xi^1, \xi^2)\right). \tag{6.1}$$

The parameter space is defined by the ranges of the parameters ξ^1 and ξ^2. It may be a square, for example, if we use parameters $\xi^1, \xi^2 \in [0, \pi]$. The physical surface is the image of the parameter space under the map $\vec{x}(\xi^1, \xi^2)$; it is a surface in target space. Alternatively, we can view the parameters ξ^1 and ξ^2 as *coordinates* on the physical surface, at least locally. The map inverse to $\vec{x}$ takes the surface to the parameter space. Locally this map is one-to-one and it assigns to each point on the surface two coordinates: the values of the parameters ξ^1 and ξ^2.

We want to calculate the area of a small element of the target space surface. Let us start by looking at an infinitesimal rectangle on the parameter space. Denote the sides of the rectangle by $d\xi^1$ and $d\xi^2$. We want to find dA, the area of the image of this little rectangle in the target space. As shown in Figure 6.3, this is the area of the actual piece of surface that corresponds to the infinitesimal rectangle on parameter space.

Of course, there is no reason why that infinitesimal area element in target space should be a rectangle. In general, it is a parallelogram. Let us call the sides of this parallelogram $d\vec{v}_1$ and $d\vec{v}_2$. They are the images under the map $\vec{x}$ of the vectors $(d\xi^1, 0)$ and $(0, d\xi^2)$, respectively. We can write them as

$$d\vec{v}_1 = \frac{\partial \vec{x}}{\partial \xi^1}\, d\xi^1\,, \quad d\vec{v}_2 = \frac{\partial \vec{x}}{\partial \xi^2}\, d\xi^2\,. \tag{6.2}$$

This makes sense: $\partial\vec{x}/\partial\xi^1$, for example, represents the rate of variation of the space coordinates with respect to ξ^1. Multiplying this rate by the length $d\xi^1$ of the horizontal side of the tiny parameter-space rectangle, gives us the vector $d\vec{v}_1$ that represents this side in the target space. Now let us calculate the area dA. Using the formula for the area of a parallelogram,

$$dA = |d\vec{v}_1||d\vec{v}_2||\sin\theta| = |d\vec{v}_1||d\vec{v}_2|\sqrt{1-\cos^2\theta}$$
$$= \sqrt{|d\vec{v}_1|^2|d\vec{v}_2|^2 - |d\vec{v}_1|^2|d\vec{v}_2|^2\cos^2\theta}, \tag{6.3}$$

where θ is the angle between the vectors $d\vec{v}_1$ and $d\vec{v}_2$. In terms of spatial dot products, we have

$$dA = \sqrt{(d\vec{v}_1 \cdot d\vec{v}_1)(d\vec{v}_2 \cdot d\vec{v}_2) - (d\vec{v}_1 \cdot d\vec{v}_2)^2}\,. \tag{6.4}$$

Finally, using (6.2),

$$dA = d\xi^1 d\xi^2 \sqrt{\Big(\frac{\partial\vec{x}}{\partial\xi^1}\cdot\frac{\partial\vec{x}}{\partial\xi^1}\Big)\Big(\frac{\partial\vec{x}}{\partial\xi^2}\cdot\frac{\partial\vec{x}}{\partial\xi^2}\Big) - \Big(\frac{\partial\vec{x}}{\partial\xi^1}\cdot\frac{\partial\vec{x}}{\partial\xi^2}\Big)^2}. \tag{6.5}$$

This is the general expression for the area element of a parameterized spatial surface. The full area functional A is given by

$$A = \int d\xi^1 d\xi^2 \sqrt{\Big(\frac{\partial\vec{x}}{\partial\xi^1}\cdot\frac{\partial\vec{x}}{\partial\xi^1}\Big)\Big(\frac{\partial\vec{x}}{\partial\xi^2}\cdot\frac{\partial\vec{x}}{\partial\xi^2}\Big) - \Big(\frac{\partial\vec{x}}{\partial\xi^1}\cdot\frac{\partial\vec{x}}{\partial\xi^2}\Big)^2}. \tag{6.6}$$

The integral extends over the relevant ranges of the parameters ξ^1 and ξ^2. The solution of a minimal area problem for a spatial surface is the function $\vec{x}\,(\xi^1, \xi^2)$ that minimizes the functional A.

6.2 Reparameterization invariance of the area

As we have seen, the parameterization of a surface allows us to write the area element in an explicit form. The area of the surface, or even more, the area of any piece of the surface, should be independent of the parameterization chosen to calculate it. This is what we mean when we say that the area must be reparameterization invariant.

Because we will soon equate the relativistic string action to some notion of proper area, it, too, will be reparameterization invariant. This means that we will be free to choose the most useful parameterization without changing the underlying physics. A good choice of parameterization will enable us to solve the equations of motion of the relativistic string in an elegant way.

Reparameterization invariance is thus an important concept so it should be understood thoroughly. To this end we will try to make it manifest in our formulae. The aim of the following analysis is to show how this can be done.

Let us begin by asking: is the area functional A in (6.6) reparameterization invariant? We would certainly hope it is. In fact, at first glance it appears to be manifestly reparameterization invariant. After all, if one reparameterizes the surface with $\tilde{\xi}^1(\xi^1)$ and $\tilde{\xi}^2(\xi^2)$, then all of the derivatives introduced by the chain rule cancel appropriately.

Quick calculation 6.1 Verify the above statement. That is, show that (6.6), written fully with tilde parameters $(\tilde{\xi}^1, \tilde{\xi}^2)$, equals (6.6) when $\tilde{\xi}^1 = \tilde{\xi}^1(\xi^1)$ and $\tilde{\xi}^2 = \tilde{\xi}^2(\xi^2)$.

The above reparameterization, however, is not completely general for it fails to mix the ξ^1 and ξ^2 coordinates. Suppose, instead, that we make a reparameterization $\tilde{\xi}^1(\xi^1, \xi^2)$ and $\tilde{\xi}^2(\xi^1, \xi^2)$. This time we can verify, using a somewhat laborious computation, that (6.6) is invariant under such a reparameterization. But the invariance is no longer intuitively clear. To make the reparameterization invariance of (6.6) manifest we will have to rewrite the area functional in a different way.

We begin by observing how the measure of integration transforms. The change-of-variable theorem from calculus tells us that

$$d\xi^1 d\xi^2 = \left|\det\left(\frac{\partial \xi^i}{\partial \tilde{\xi}^j}\right)\right| d\tilde{\xi}^1 d\tilde{\xi}^2 = |\det M|\, d\tilde{\xi}^1 d\tilde{\xi}^2 \,, \tag{6.7}$$

where $M = [M_{ij}]$ is the matrix defined by $M_{ij} = \partial \xi^i / \partial \tilde{\xi}^j$. Similarly,

$$d\tilde{\xi}^1 d\tilde{\xi}^2 = \left|\det\left(\frac{\partial \tilde{\xi}^i}{\partial \xi^j}\right)\right| d\xi^1 d\xi^2 = |\det \tilde{M}|\, d\xi^1 d\xi^2 \,, \tag{6.8}$$

where $\tilde{M} = [\tilde{M}_{ij}]$ is the matrix defined by $\tilde{M}_{ij} = \partial \tilde{\xi}^i / \partial \xi^j$. Combining equations (6.7) and (6.8), we see that

$$|\det M|\,|\det \tilde{M}| = 1 \,. \tag{6.9}$$

Let us now consider a target space surface $\mathcal{S}$ described by the mapping functions $\vec{x}(\xi^1, \xi^2)$. Given a vector $d\vec{x}$ tangent to the surface, let ds denote its length. Then we can write

$$ds^2 \equiv (ds)^2 = d\vec{x} \cdot d\vec{x}. \tag{6.10}$$

For surfaces in space, as we are considering now, it is *not* customary to include a minus sign in front of ds^2 (compare with (2.21)). The vector $d\vec{x}$ can be expressed in terms of partial derivatives and the differentials $d\xi^1$, $d\xi^2$:

$$d\vec{x} = \frac{\partial \vec{x}}{\partial \xi^1} d\xi^1 + \frac{\partial \vec{x}}{\partial \xi^2} d\xi^2 = \frac{\partial \vec{x}}{\partial \xi^i}\, d\xi^i \,. \tag{6.11}$$

The repeated index i is summed over its possible values 1 and 2. Back in (6.10),

$$ds^2 = \left(\frac{\partial \vec{x}}{\partial \xi^i} d\xi^i\right) \cdot \left(\frac{\partial \vec{x}}{\partial \xi^j} d\xi^j\right) = \frac{\partial \vec{x}}{\partial \xi^i} \cdot \frac{\partial \vec{x}}{\partial \xi^j}\, d\xi^i d\xi^j \,. \tag{6.12}$$

This can be neatly summarized as

$$ds^2 = g_{ij}(\xi)\, d\xi^i d\xi^j \,, \tag{6.13}$$

where $g_{ij}(\xi)$ is defined as

$$g_{ij}(\xi) \equiv \frac{\partial \vec{x}}{\partial \xi^i} \cdot \frac{\partial \vec{x}}{\partial \xi^j} \,. \tag{6.14}$$

The quantity $g_{ij}(\xi)$ is known as the *induced metric on* $\mathcal{S}$. It is called a metric because (6.13) takes, up to a sign, the form of equation (3.78), where we introduced the general concept of a metric. It is a metric on $\mathcal{S}$ because, with ξ^i playing the role of coordinates on $\mathcal{S}$, equation (6.13) determines distances on $\mathcal{S}$. It is said to be induced because it uses the metric on the ambient space in which $\mathcal{S}$ *lives* to determine distances on $\mathcal{S}$. Indeed, the dot product which

appears in (6.14) is to be performed in the space where $\mathcal{S}$ lives and therefore presupposes that a metric exists on that space. We only have two parameters ξ^1 and ξ^2, so the full matrix g_{ij} takes the form:

$$g_{ij} = \begin{pmatrix} \frac{\partial \vec{x}}{\partial \xi^1} \cdot \frac{\partial \vec{x}}{\partial \xi^1} & \frac{\partial \vec{x}}{\partial \xi^1} \cdot \frac{\partial \vec{x}}{\partial \xi^2} \\ \frac{\partial \vec{x}}{\partial \xi^2} \cdot \frac{\partial \vec{x}}{\partial \xi^1} & \frac{\partial \vec{x}}{\partial \xi^2} \cdot \frac{\partial \vec{x}}{\partial \xi^2} \end{pmatrix} . \tag{6.15}$$

Now we see something truly nice! The determinant of g_{ij} is precisely the quantity which appears under the square root in (6.6). Letting

$$g \equiv \det(g_{ij}) , \tag{6.16}$$

we can write

$$A = \int d\xi^1 d\xi^2 \sqrt{g} . \tag{6.17}$$

This is an elegant formula for the area in terms of the determinant of the induced metric. Instead of trying to understand the reparameterization invariance of (6.6), we now focus on the equivalent but simpler expression (6.17).

We are now in position to understand the invariance of the area in terms of the transformation properties of the metric g_{ij}. The key to this lies in equation (6.13). The length-squared ds^2 is a geometrical property of the vector $d\vec{x}$ that must not depend upon the particular parameterization used to calculate it. For another set of parameters $\tilde{\xi}$ and metric $\tilde{g}(\tilde{\xi})$, the following equality must therefore hold:

$$g_{ij}(\xi)\, d\xi^i d\xi^j = \tilde{g}_{pq}(\tilde{\xi})\, d\tilde{\xi}^p d\tilde{\xi}^q . \tag{6.18}$$

Making use of the chain rule to express the differentials $d\tilde{\xi}$ in terms of differentials $d\xi$,

$$g_{ij}(\xi)\, d\xi^i d\xi^j = \tilde{g}_{pq}(\tilde{\xi}) \frac{\partial \tilde{\xi}^p}{\partial \xi^i} \frac{\partial \tilde{\xi}^q}{\partial \xi^j} d\xi^i d\xi^j . \tag{6.19}$$

Since this result holds for any choice of differentials $d\xi$, we find a relation between the metric in ξ and $\tilde{\xi}$ coordinates:

$$g_{ij}(\xi) = \tilde{g}_{pq}(\tilde{\xi}) \frac{\partial \tilde{\xi}^p}{\partial \xi^i} \frac{\partial \tilde{\xi}^q}{\partial \xi^j} . \tag{6.20}$$

Making use of the definition of $\tilde{M}$ below (6.8), we rewrite the above equation as

$$g_{ij}(\xi) = \tilde{g}_{pq}\, \tilde{M}_{pi}\, \tilde{M}_{qj} = (\tilde{M}^{\mathrm{T}})_{ip}\, \tilde{g}_{pq}\, \tilde{M}_{qj} . \tag{6.21}$$

In matrix notation, the right-hand side is the product of three matrices. Taking the determinant and using the notation in (6.16) gives

$$g = (\det \tilde{M}^{\mathrm{T}})\, \tilde{g}\, (\det \tilde{M}) = \tilde{g} (\det \tilde{M})^2 . \tag{6.22}$$

Taking a square root

$$\sqrt{g} = \sqrt{\tilde{g}}\,|\det \tilde{M}|\,, \tag{6.23}$$

we obtain the transformation property for the square root of the determinant of the metric.

We are finally ready to appreciate the reparameterization invariance of (6.17). Making use of (6.7), (6.23), and (6.9) we have

$$\int d\xi^1 d\xi^2 \sqrt{g} = \int d\tilde{\xi}^1 d\tilde{\xi}^2 |\det M| \sqrt{\tilde{g}}\,|\det \tilde{M}| = \int d\tilde{\xi}^1 d\tilde{\xi}^2 \sqrt{\tilde{g}}\,, \tag{6.24}$$

which proves the reparameterization invariance of the area functional. To the trained eye the area formula in (6.17) is *manifestly* reparameterization invariant. That is, once you know how metrics transform, the invariance is reasonably simple to establish. No cumbersome calculation is necessary.

Quick calculation 6.2 Consider the equation $\partial\xi^i/\partial\xi^j = \delta^i_j$ and use the chain rule to show the matrix property

$$M\tilde{M} = 1\,. \tag{6.25}$$

Show that $\tilde{M}M = 1$ holds as well. Finally, note that $\det M \det \tilde{M} = 1$, a result stronger than the one we proved in (6.9).

6.3 Area functional for spacetime surfaces

Let us now move to our case of interest, the case of surfaces in *spacetime*. These surfaces are obtained by representing in spacetime the history of strings, in the same way as a spacetime world-line is obtained by representing the history of a particle. For the case of strings, we obtain a two-dimensional surface called the world-sheet of the string. Spacetime surfaces, such as string world-sheets, are not all that different from the spatial surfaces we considered in the previous section. They are two-dimensional and require two parameters. Instead of calling the parameters ξ^1 and ξ^2, we give them special names: τ and σ.

Given our usual spacetime coordinates $x^\mu = (x^0, x^1, \ldots, x^d)$, the surface is described by the mapping functions

$$x^\mu(\tau, \sigma)\,, \tag{6.26}$$

which take some region of the (τ, σ) parameter space into spacetime. Following a standard convention in string theory, we change the notation slightly. We will denote the above mapping functions with the capitalized symbols

$$X^\mu(\tau, \sigma)\,. \tag{6.27}$$

We are not changing the meaning of the functions. Given a fixed point (τ, σ) in the parameter space, this point is mapped to a point with spacetime coordinates

$$(X^0(\tau, \sigma),\ X^1(\tau, \sigma), \ldots,\ X^d(\tau, \sigma)). \tag{6.28}$$

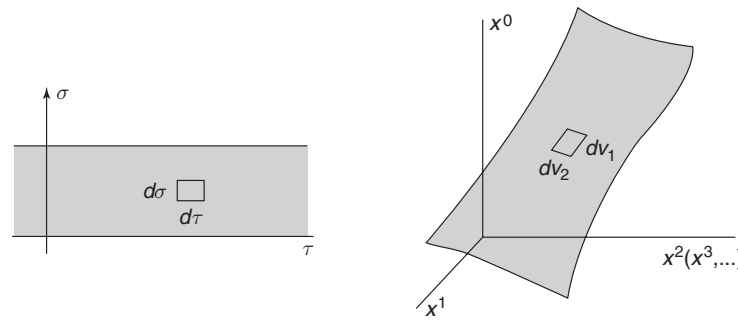

Fig. 6.4 **Left: the parameter space (τ, σ), with a little square selected. Right: the surface in target spacetime with the image of the little square, a parallelogram whose sides are the vectors dv_1^μ and dv_2^μ.**

Why do we capitalize the X? Suppose we used the same symbol to denote spacetime coordinates and mapping functions. Then we could still distinguish between them by writing x^μ or $x^\mu(\tau, \sigma)$, but we would not have the luxury of dropping the (τ, σ) arguments. On the other hand, with X^μ we can drop the (τ, σ) arguments and still know that we are talking about the mapping functions of the string. We will call X^μ the *string coordinates*.

As before, the parameters τ and σ can be viewed as coordinates on the world-sheet, at least locally. The map inverse to X^μ takes the world-sheet to the parameter space, and locally it assigns to each point on the surface two coordinates: the values of the parameters τ and σ. Introducing some potential for confusion, physicists also use the term world-sheet to denote the two-dimensional parameter space whose image under X^μ gives us the... world-sheet! Unless explicitly stated, we will reserve the use of the term world-sheet for the spacetime surface. In Figure 6.4 we consider an open string: to the left, you see the parameter space surface and, to the right, you see the spacetime surface. In this parameter space, σ ranges over a finite interval, while τ may extend from minus infinity to plus infinity. The parameter τ is roughly related to time on the strings – much more on this later – and the parameter σ is roughly related to positions along the strings. The world-lines of the string endpoints have constant σ, so they are parameterized by τ. As τ flows, time must flow. Thus, at least at the endpoints

$$\left.\frac{\partial X^0}{\partial \tau}\right|_{\text{endpoint}} \neq 0\,. \tag{6.29}$$

We will assume that this also holds for other values of σ.

To find the area element we proceed as in the case of the spatial surface, this time using relativistic notation. The situation is illustrated in Figure 6.4. A little rectangle of sides $d\tau$ and $d\sigma$ in parameter-space becomes a quadrilateral area element in spacetime. This quadrilateral is spanned by the vectors dv_1^μ and dv_2^μ. Furthermore,

$$dv_1^\mu = \frac{\partial X^\mu}{\partial \tau} d\tau\,, \quad dv_2^\mu = \frac{\partial X^\mu}{\partial \sigma} d\sigma, \tag{6.30}$$

which are analogous to our earlier spatial formulae (6.2). We can now use the analog of (6.4) as a candidate for the area element dA:

$$dA \stackrel{?}{=} \sqrt{(dv_1 \cdot dv_1)(dv_2 \cdot dv_2) - (dv_1 \cdot dv_2)^2}\,, \tag{6.31}$$

where the dot is the relativistic dot product. Using this dot product guarantees that the area element is Lorentz invariant: it is a proper area element. We wrote a question mark on top of the equal sign because there is one problem. Even though this is not obvious to us yet, the sign of the object under the square root is negative. To be able to take the square root we must exchange the two terms under the square root. This change of sign has no effect on the Lorentz invariance. Doing this, and using (6.30), we find that the proper area is given as

$$A = \int d\tau d\sigma \sqrt{\left(\frac{\partial X^\mu}{\partial \tau}\frac{\partial X_\mu}{\partial \sigma}\right)^2 - \left(\frac{\partial X^\mu}{\partial \tau}\frac{\partial X_\mu}{\partial \tau}\right)\left(\frac{\partial X^\nu}{\partial \sigma}\frac{\partial X_\nu}{\partial \sigma}\right)}\,. \tag{6.32}$$

Using the relativistic dot product notation,

$$A = \int d\tau d\sigma \sqrt{\left(\frac{\partial X}{\partial \tau}\cdot\frac{\partial X}{\partial \sigma}\right)^2 - \left(\frac{\partial X}{\partial \tau}\right)^2\left(\frac{\partial X}{\partial \sigma}\right)^2}\,. \tag{6.33}$$

To understand why the above sign is correct we must convince ourselves that the expression under the square root is greater than or equal to zero at any point on the world-sheet of a string.

What characterizes locally the spacetime surface traced by a string? The answer is quite interesting. Consider a point on the world-sheet and the set of all vectors tangent to the surface at that point. These vectors form a two-dimensional vector space. We claim that in this vector space there is a basis made by two vectors, one of which is spacelike and one of which is timelike. This implies that at each point on the world-sheet there are both timelike and spacelike tangent directions. There is a small caveat: on each fixed-time string there can be a finite set of exceptional points where the tangents to the world-sheet do not include a timelike vector. At those points, as we will see, the string moves with the speed of light.

The string we are trying to define cannot have *finite* size pieces moving with the speed of light. In such case, a string would contain a continuous set of points, not just a finite set, where the tangents to the world-sheet do not include timelike vectors. Imagine, for example, a straight piece of string along the x axis moving with the speed of light in the y direction. At any point on this string all vectors tangent to the world-sheet are spacelike, except for null vectors in one particular direction. To see this consider Figure 6.5 where we show this string at various closely separated times. Tangent vectors at P are well approximated by vectors that joint the event P to a nearby event P' on the *world-sheet*. Since P and P' are points on the same string at different times, the spatial part of the tangent vector appears in the figure as the arrow joining P to P'. Consider all arrows joining P to points on a semicircle around P. Any world-sheet tangent direction at P is represented by one of these arrows. The world-sheet tangent vector associated with PQ is clearly spacelike, since P and Q occur at the same time. The typical tangent, that associated with the arrow PR is still spacelike: in the elapsed time P can get to $\bar{R}$ moving at the speed of light, but to

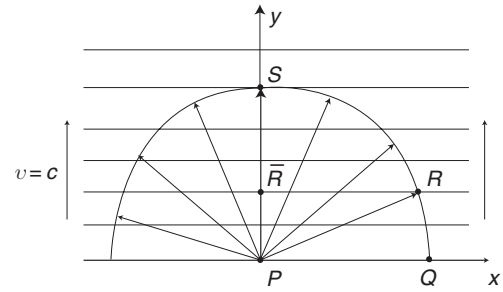

Fig. 6.5 **A string along the *x* direction moving with the velocity of light along the *y* direction. This is not allowed motion. All tangent vectors to the world-sheet at any point *P* on the string are either spacelike or null.**

get to R it must move faster than light. All tangent directions are spacelike, except for that associated with PS, which is null. This shows that there is no timelike tangent at P. Since P is a generic point along the original piece of string, nowhere on this piece of string is there a timelike vector tangent to the world-sheet.

At any point on a world-sheet a spacelike direction exists and is easy to visualize: if you take a photograph of the string at some time, every tangent vector along the length of the string points in a spacelike direction. Indeed, in your frame, the events defining the string are simultaneous but spatially separated.

To appreciate the need for a timelike vector at regular points on the world-sheet, consider first the world-line of a point particle. The tangent vector to the world-line is timelike. At each point on the world-line this tangent vector can be used to describe an instantaneous Lorentz observer who sees the particle at rest. A spacelike tangent vector to the world-line is unphysical: it describes a particle moving faster than the speed of light. Strings are a little more subtle since there is no way to tell how individual points on them move. As we shall make abundantly clear, the string is not made of constituents whose position we can keep track of (exception: one can keep track of the endpoints of an open string). Still, a timelike tangent to the world-sheet at a given point of a string allows us to describe an instantaneous Lorentz observer who sees the point at rest. If there is a timelike tangent at a point, there are many, by continuity. Each of these timelike tangents defines a different instantaneous Lorentz observer who sees the point at rest. This is consistent with our inability to track unambiguously the motion of points on a string. If we watch a string at two closely separated times we cannot tell which point went where, but for each point p on the final string we must be able to find some point p' on the initial string that could reach p moving with speed less than or at most equal to c.

The existence of both timelike directions and spacelike directions at any regular point on the world-sheet is our criterion for physical motion. It guarantees that equation (6.33) makes sense.

Claim: At any point P on the world-sheet where there is both a timelike direction and a spacelike direction, the quantity under the square root in (6.33) is positive:

$$\left(\frac{\partial X}{\partial \tau}\cdot\frac{\partial X}{\partial \sigma}\right)^2-\left(\frac{\partial X}{\partial \sigma}\right)^2\left(\frac{\partial X}{\partial \tau}\right)^2>0\,. \tag{6.34}$$

Proof: Consider the set of tangent vectors $v^\mu(\lambda)$ at P obtained as:

$$v^\mu(\lambda) = \frac{\partial X^\mu}{\partial \tau} + \lambda \frac{\partial X^\mu}{\partial \sigma}\,, \tag{6.35}$$

where $\lambda \in (-\infty, \infty)$ is a parameter. Since $\partial X^\mu/\partial\tau$ and $\partial X^\mu/\partial\sigma$ are linearly independent tangent vectors, when we vary λ we get, up to constant scalings, all tangent vectors at P, including $\partial X^\mu/\partial\sigma$, which is obtained in the limit $\lambda \to \infty$ (Figure 6.6). Constant scalings of a vector do not matter to decide if a vector is timelike or spacelike. To determine if $v^\mu(\lambda)$ is timelike or spacelike, we consider its square:

$$v^2(\lambda) = v^\mu(\lambda) v_\mu(\lambda) = \lambda^2 \Big(\frac{\partial X}{\partial \sigma}\Big)^2 + 2\lambda \Big(\frac{\partial X}{\partial \tau} \cdot \frac{\partial X}{\partial \sigma}\Big) + \Big(\frac{\partial X}{\partial \tau}\Big)^2 . \tag{6.36}$$

The dot products appearing on the final right-hand side are just numbers, so we have a quadratic polynomial in λ. To have both timelike and spacelike tangent vectors at P, $v^2(\lambda)$ must take both negative and positive values as we vary λ. In other words, the equation $v^2(\lambda) = 0$ must have two real roots. For this to happen, the discriminant of the quadratic equation $v^2(\lambda) = 0$ must be positive. From (6.36) we see that this requires

$$\Big(\frac{\partial X}{\partial \tau} \cdot \frac{\partial X}{\partial \sigma}\Big)^2 - \Big(\frac{\partial X}{\partial \sigma}\Big)^2 \Big(\frac{\partial X}{\partial \tau}\Big)^2 > 0\,, \tag{6.37}$$

which is precisely the condition (6.34) we set out to prove!

Since we always have spacelike tangents, the plot of $v^2(\lambda)$ in Figure 6.6 must include a region where $v^2(\lambda) > 0$. Consider a point P on the world-sheet where all tangent directions are spacelike with the exception of one that is null. Then $v^2(\lambda) > 0$, except for one value of λ where v^2 vanishes. The equation $v^2(\lambda) = 0$ must have a single root and the associated discriminant is zero. It follows that the quantity under the square root in the action (6.33) is zero at P. Any possible motion of the string at P must be associated with a world-sheet tangent at P. Since motion along spacelike directions is unphysical, only the null vector provides an acceptable answer: the string is moving with the speed of light at P.

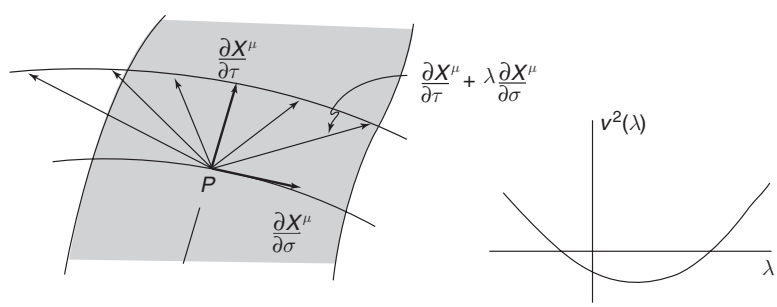

Fig. 6.6 Left: a set of tangent vectors $v(\lambda)$ at a point P on the world-sheet. Right: a plot of $v^2(\lambda)$ as a function of λ. The vector $v(\lambda)$ may be spacelike or timelike depending on the value of λ.

6.4 The Nambu–Goto string action

Now that we are sure that the proper area functional in (6.33) is correctly defined, we can introduce the action for the relativistic string. This action is proportional to the proper area of the world-sheet. To have the units of action we must multiply the area functional by some suitable constants.

The area functional in (6.33) has units of length-squared, as it must be. This is because X^μ has units of length, and each term under the square root has four X. The units of τ and σ cancel out. Each term in the square root has two σ derivatives and two τ derivatives. Their units cancel against the units of the differentials. Nevertheless, we will take σ to have units of length and τ to have units of time. We do this anticipating a relation between τ and time and between σ and positions on strings. To summarize:

$$[\tau] = T\,, \quad [\sigma] = L\,, \quad [X^\mu] = L\,, \quad [A] = L^2\,. \tag{6.38}$$

Since S must have units of ML^2/T and A has units of L^2, we must multiply the proper area by a quantity with units of M/T. The string tension T_0 has units of force, and force divided by velocity has the desired units of M/T. We can therefore multiply the proper area by T_0/c to get a quantity with the units of action. Making use of (6.33) we set the string action equal to

$$S = -\frac{T_0}{c}\int_{\tau_i}^{\tau_f} d\tau \int_0^{\sigma_1} d\sigma \sqrt{(\dot{X}\cdot X')^2 - (\dot{X})^2 (X')^2}\,. \tag{6.39}$$

Here $\sigma_1 > 0$ is some constant, and we have introduced some notation for derivatives:

$$\dot{X}^\mu \equiv \frac{\partial X^\mu}{\partial \tau}\,, \quad X^{\mu\prime} \equiv \frac{\partial X^\mu}{\partial \sigma}\,. \tag{6.40}$$

Of course, we have not yet confirmed that the symbol T_0 in the string action has the precise interpretation of tension, but we will do so in Section 6.7. We will also confirm there that the overall negative sign which multiplies the action is correct. The action S is the Nambu–Goto action for the relativistic string.

It is crucial that this action be reparameterization invariant. We can proceed just as we did with spatial surfaces to write the Nambu–Goto action in a manifestly reparameterization invariant way. In this case we have

$$-ds^2 = dX^\mu dX_\mu = \eta_{\mu\nu}\, dX^\mu dX^\nu = \eta_{\mu\nu}\frac{\partial X^\mu}{\partial \xi^\alpha}\frac{\partial X^\nu}{\partial \xi^\beta}\, d\xi^\alpha d\xi^\beta\,. \tag{6.41}$$

Here $\eta_{\mu\nu}$ is the target-space Minkowski metric. The indices α and β run over two values, 1 and 2, and we have taken $\xi^1 = \tau$, $\xi^2 = \sigma$. Just as we did for spatial surfaces, we define an induced metric $\gamma_{\alpha\beta}$ on the world-sheet:

$$\gamma_{\alpha\beta} \equiv \eta_{\mu\nu}\frac{\partial X^\mu}{\partial \xi^\alpha}\frac{\partial X^\nu}{\partial \xi^\beta} = \frac{\partial X}{\partial \xi^\alpha}\cdot\frac{\partial X}{\partial \xi^\beta}\,. \tag{6.42}$$

More explicitly, the 2-by-2 matrix $\gamma_{\alpha\beta}$ is

$$\gamma_{\alpha\beta} = \begin{bmatrix} (\dot{X})^2 & \dot{X}\cdot X' \\ \dot{X}\cdot X' & (X')^2 \end{bmatrix}. \tag{6.43}$$

With the help of this metric we can write the Nambu–Goto action in the manifestly reparameterization invariant form

$$S = -\frac{T_0}{c}\int d\tau d\sigma \sqrt{-\gamma}\,, \quad \gamma = \det(\gamma_{\alpha\beta})\,. \tag{6.44}$$

The analysis in Section 6.2 of reparameterization invariance for spatial surfaces holds, without change, in the present case. Not only is the action (6.44) manifestly reparameterization invariant, it is also more compact than (6.39). In this form, one can readily generalize the Nambu–Goto action to describe the dynamics of objects that have more dimensions than strings. An action of this kind is useful as a first approximation to the dynamics of D-branes.

6.5 Equations of motion, boundary conditions, and D-branes

In this section we will obtain the equations of motion that follow by variation of the string action. In doing so we will also have an opportunity to discuss the various boundary conditions that can be imposed on the ends of open strings. Dirichlet boundary conditions will be interpreted to arise owing to the existence of D-branes.

Let us begin by writing the Nambu–Goto action (6.39) as the double integral of a Lagrangian density $\mathcal{L}$:

$$S = \int_{\tau_i}^{\tau_f} d\tau L = \int_{\tau_i}^{\tau_f} d\tau \int_0^{\sigma_1} d\sigma\ \mathcal{L}(\dot{X}^\mu, X^{\mu\prime})\,, \tag{6.45}$$

where $\mathcal{L}$ is given by

$$\mathcal{L}(\dot{X}^\mu, X^{\mu\prime}) = -\frac{T_0}{c}\sqrt{(\dot{X}\cdot X')^2 - (\dot{X})^2(X')^2}\,. \tag{6.46}$$

We can obtain the equations of motion for the relativistic string by setting the variation of the action (6.45) equal to zero. The variation is simply

$$\delta S = \int_{\tau_i}^{\tau_f} d\tau \int_0^{\sigma_1} d\sigma \left[\frac{\partial\mathcal{L}}{\partial\dot{X}^\mu}\frac{\partial(\delta X^\mu)}{\partial\tau} + \frac{\partial\mathcal{L}}{\partial X^{\mu\prime}}\frac{\partial(\delta X^\mu)}{\partial\sigma}\right], \tag{6.47}$$

where we have used

$$\delta\dot{X}^\mu = \delta\Big(\frac{\partial X^\mu}{\partial\tau}\Big) = \frac{\partial(\delta X^\mu)}{\partial\tau}\,, \tag{6.48}$$

and an analogous equation for $\delta X^{\mu\prime}$.

The quantities $\partial\mathcal{L}/\partial\dot{X}^\mu$ and $\partial\mathcal{L}/\partial X'^\mu$ will appear frequently throughout the remainder of our discussion, so it is useful to introduce new symbols for them. This is just what we did when we studied the nonrelativistic string in Section 4.6. This time we find

$$\mathcal{P}^\tau_\mu \equiv \frac{\partial\mathcal{L}}{\partial\dot{X}^\mu} = -\frac{T_0}{c}\frac{(\dot{X}\cdot X')X'_\mu - (X')^2\dot{X}_\mu}{\sqrt{(\dot{X}\cdot X')^2 - (\dot{X})^2(X')^2}}\,, \tag{6.49}$$

$$\mathcal{P}^\sigma_\mu \equiv \frac{\partial\mathcal{L}}{\partial X^{\mu\prime}} = -\frac{T_0}{c}\frac{(\dot{X}\cdot X')\dot{X}_\mu - (\dot{X})^2X'_\mu}{\sqrt{(\dot{X}\cdot X')^2 - (\dot{X})^2(X')^2}}\,. \tag{6.50}$$

Quick calculation 6.3 Verify equations (6.49) and (6.50).

Using this notation, the variation δS in (6.47) becomes

$$\delta S = \int_{\tau_i}^{\tau_f} d\tau \int_0^{\sigma_1} d\sigma \left[\frac{\partial}{\partial\tau}(\delta X^\mu\mathcal{P}^\tau_\mu) + \frac{\partial}{\partial\sigma}\left(\delta X^\mu\mathcal{P}^\sigma_\mu\right) - \delta X^\mu\Big(\frac{\partial\mathcal{P}^\tau_\mu}{\partial\tau} + \frac{\partial\mathcal{P}^\sigma_\mu}{\partial\sigma}\Big)\right]. \tag{6.51}$$

The first term on the right-hand side, being a full derivative in τ, will contribute terms proportional to $\delta X^\mu(\tau_f,\sigma)$ and $\delta X^\mu(\tau_i,\sigma)$. Since the flow of τ implies the flow of time, we can imagine specifying the initial and final states of the string, and we restrict ourselves to variations for which $\delta X^\mu(\tau_f,\sigma) = \delta X^\mu(\tau_i,\sigma) = 0$. We will always assume such variations, so we can forget about these terms. The variation then becomes

$$\delta S = \int_{\tau_i}^{\tau_f} d\tau\left[\delta X^\mu\,\mathcal{P}^\sigma_\mu\right]_0^{\sigma_1} - \int_{\tau_i}^{\tau_f} d\tau\int_0^{\sigma_1} d\sigma\,\delta X^\mu\left(\frac{\partial\mathcal{P}^\tau_\mu}{\partial\tau} + \frac{\partial\mathcal{P}^\sigma_\mu}{\partial\sigma}\right). \tag{6.52}$$

Since the second term on the right-hand side must vanish for all variations δX^μ of the motion, we set

$$\frac{\partial\mathcal{P}^\tau_\mu}{\partial\tau} + \frac{\partial\mathcal{P}^\sigma_\mu}{\partial\sigma} = 0\,. \tag{6.53}$$

This is the equation of motion for the relativistic string, open or closed. A quick glance at definitions (6.49) and (6.50) shows that this equation is incredibly complicated. The key to its solution will lie in the reparameterization invariance of the Nambu–Goto action. Choosing a clever parameterization will simplify our work enormously.

The first term on the right-hand side of (6.52) has to do with the string endpoints. It is, in fact, a collection of terms that includes two terms for each value of the index μ. More explicitly, the list is

$$\int_{\tau_i}^{\tau_f} d\tau \Big(\delta X^0(\tau,\sigma_1)\,\mathcal{P}_0^\sigma(\tau,\sigma_1) - \delta X^0(\tau,0)\mathcal{P}_0^\sigma(\tau,0) + \delta X^1(\tau,\sigma_1)\mathcal{P}_1^\sigma(\tau,\sigma_1) - \delta X^1(\tau,0)\mathcal{P}_1^\sigma(\tau,0) \quad \vdots \quad \vdots \quad \vdots \quad \vdots + \delta X^d(\tau,\sigma_1)\mathcal{P}_d^\sigma(\tau,\sigma_1) - \delta X^d(\tau,0)\mathcal{P}_d^\sigma(\tau,0) \Big) . \tag{6.54}$$

We need a boundary condition for *each* term in the above list. This is a total of $2D = 2(d+1)$ boundary conditions.

Let us focus on a single term, that is, we fix μ and select one endpoint. Let σ_* denote the σ coordinate of an endpoint; σ_* can be equal to zero or equal to σ_1. Selecting an endpoint means fixing the value of σ_*. As before, there are two natural boundary conditions that one can impose at an endpoint. The first is a Dirichlet boundary condition, in which the endpoint of the string remains fixed throughout the motion:

$$\text{Dirichlet boundary condition:} \quad \frac{\partial X^\mu}{\partial \tau}(\tau,\sigma_*) = 0\,, \quad \mu \neq 0\,. \tag{6.55}$$

Since time varies as τ varies (see (6.29)), the value $\mu = 0$ must be excluded. Dirichlet boundary conditions are only possible for space directions. Given that constancy in τ means constancy in time, equation (6.55) implies that the μ coordinate of the selected string endpoint is fixed in time. Alternatively, rather than requiring that the τ derivative vanishes, we could simply specify a constant value for $X^\mu(\tau,\sigma_*)$. If the string endpoint is fixed, the variations are set to vanish there: $\delta X^\mu(\tau\,,\sigma_*) = 0$. This guarantees that the relevant term in (6.5) vanishes.

The second possible boundary condition is a free endpoint condition:

$$\text{free endpoint condition:} \quad \mathcal{P}_\mu^\sigma(\tau,\sigma_*) = 0\,. \tag{6.56}$$

This condition, as needed, also results in the vanishing of the relevant term in (6.5). This is called a free endpoint condition because it does not impose any constraint on the variation $\delta X^\mu(\tau,\sigma_*)$ of the string coordinate at the endpoint. The endpoint is free to do whatever is needed to get the variation of the action to vanish. The free endpoint boundary condition must apply for $\mu = 0$:

$$\mathcal{P}_0^\sigma(\tau,\sigma_1) = \mathcal{P}_0^\sigma(\tau,0) = 0\,. \tag{6.57}$$

For the nonrelativistic string, the free endpoint boundary condition implies the vanishing of $\mathcal{P}^x$, which imposes a Neumann boundary condition on the string coordinate (see (4.47)).

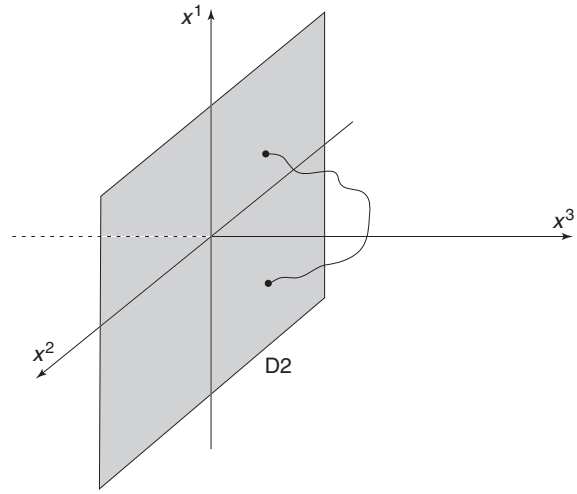

Fig. 6.7 **A D2-brane stretched over the (x^1, x^2) plane. The endpoints of the open string can move freely on the plane, but must remain attached to it. The coordinate x^3 of the endpoints must vanish at all times. This is a Dirichlet boundary condition for the string coordinate X^3.**

We will eventually understand (6.56) in terms of a Neumann boundary condition. Similarly, the Dirichlet boundary (6.55) will be shown to imply the vanishing of $\mathcal{P}^\tau_\mu$ at the string endpoints.

The boundary conditions (6.55) and (6.56) can be imposed in many possible ways. For each spatial direction, and at each endpoint, we can choose either a Dirichlet or a free-endpoint boundary condition. Since closed strings have no endpoints, they do not require boundary conditions.

Let us elaborate on the case of Dirichlet boundary conditions. It is clear from the study of nonrelativistic strings that Dirichlet boundary conditions arise if string endpoints are attached to some physical objects. Consider, for example, Figure 4.2. On the left, the string is attached to two points. On the right, the string is free to slide up and down at the endpoints; the string endpoints are forced to stay on one-dimensional lines and horizontal motion of the endpoints is forbidden. The objects on which open string endpoints must lie are characterized by their dimensionality, more precisely, by the number of spatial dimensions that they have. They are called D-branes, where the letter D stands for Dirichlet. The objects which fix the string endpoints on the left side of Figure 4.2 are zero-dimensional. They are called D0-branes. The lines which constrain the string endpoints on the right side of the figure are one dimensional. They are called D1-branes.

A Dp-brane is an object with p spatial dimensions. Since the string endpoints must lie on the Dp-brane, a set of Dirichlet boundary conditions is specified. A flat D2-brane in a three-dimensional space, for example, is specified by one condition, say $x^3 = 0$ (Figure 6.7). This means that the D2-brane extends over the (x^1, x^2) plane. The Dirichlet boundary condition applies to the string coordinate X^3, which must vanish at the string endpoints. Since the motion of open string endpoints is free along the directions of the brane, the string coordinates X^1 and X^2 satisfy free boundary conditions. When the open string endpoints

have free boundary conditions along all spatial directions, we still have a D-brane, but this time it is a *space-filling* D-brane. The D-brane extends all over space, and since open string endpoints can be anywhere on the D-brane, open string endpoints are completely free.

For (quantum) relativistic strings the consistency of Dirichlet boundary conditions allows one to discover the properties of D-branes. D-branes are physical objects that exist in a theory of strings and are not introduced by hand. D-branes need not have infinite extent nor are they necessarily hyperplanes. They have calculable energy densities and a host of remarkable properties. We will study D-branes in detail beginning in Chapter 15.

6.6 The static gauge

To make progress in understanding the action for the relativistic string, we must parameterize the world-sheet in a useful way. We are allowed to choose the parameterization freely because of the reparameterization invariance of the string action. Reparameterization invariance in string theory is analogous to gauge invariance in electrodynamics. Maxwell's equations possess a symmetry under gauge transformations that allows us to use different potentials A_μ to represent the same electromagnetic fields $\vec{E}$ and $\vec{B}$. A suitable choice of gauge helps to uncover the physics. Similarly, we may use many different grids on the world-sheet to describe the same physical motion of the string. A suitable choice of grid can make this task much easier. A good choice of parameterization was useful even for the relativistic point particle – its equation of motion is simplest when the trajectory is parameterized by proper time.

In this section, we will discuss only a partial parameterization on the world-sheet. We will fix the lines of constant τ by relating τ to the time coordinate $X^0 = ct$ in some chosen Lorentz frame.

Consider the constant time hyperplane $t = t_0$ in the target space (Figure 6.8). This plane will intersect the world-sheet along a curve – *the string at time* t_0 according to observers in our chosen Lorentz frame. We declare this curve to be a curve of constant τ; in fact, we declare it to be the curve $\tau = t_0$. Extending this definition to all times t, we declare that for any point Q on the world-sheet

$$\tau(Q) = t(Q). \tag{6.58}$$

This choice of τ parameterization is called the *static gauge* because lines of constant τ are "static strings" in the chosen Lorentz frame.

We will not try to make a sophisticated choice of σ at this time. For an open string, we will choose one edge of the world-sheet to be the curve $\sigma = 0$ and the other edge to be the curve $\sigma = \sigma_1$:

$$\sigma \in [0, \sigma_1], \quad \text{for an open string.} \tag{6.59}$$

We draw lines of constant σ on the surface quite arbitrarily, provided, of course, that constant σ lines vary smoothly, do not intersect, and are consistent with the two curves which

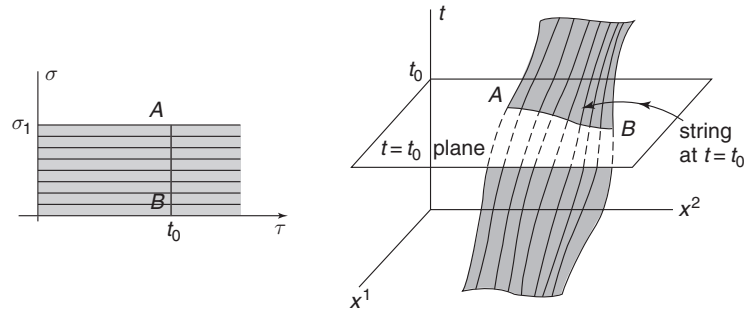

Fig. 6.8 **Left: the parameter space strip for an open string. The vertical segment *AB* is the line $\tau = t_0$. Right: the open string world-sheet in target space. The string at time $t = t_0$ is the intersection of the world-sheet with the hyperplane $t = t_0$. In the static gauge, the string at time $t = t_0$ is the image of the $\tau = t_0$ segment *AB*.**

are the boundary of the world-sheet (Figure 6.8). Drawing lines of constant σ is equivalent to giving an explicit σ parameterization to all the strings. For closed strings the same ideas apply, but there is a significant proviso: there must be an identification in the (τ, σ) parameter space. The σ direction must be made into a circle, making the (τ, σ) parameter space into a cylinder. This is needed because the closed string world-sheet is topologically a cylinder. Letting σ_c denote the circumference of the σ circle, the identification is

$$(\tau, \sigma) \sim (\tau, \sigma + \sigma_c)\,. \tag{6.60}$$

Points that are identified by this relation on the parameter space map to the same point on the closed string world-sheet. The closed strings can be parameterized using any σ interval of length σ_c, for example

$$\sigma \in [0, \sigma_c]\,, \quad \text{for a closed string.} \tag{6.61}$$

Let us now explore some implications of our choice of τ. We can write (6.58) as

$$X^0(\tau, \sigma) \equiv c\,t(\tau, \sigma) = c\,\tau\,, \tag{6.62}$$

or simply

$$\tau = t\,. \tag{6.63}$$

We can thus describe the collection of string coordinates X^μ as

$$X^\mu(\tau, \sigma) = X^\mu(t, \sigma) = \{c\,t, \vec{X}(t, \sigma)\}\,, \tag{6.64}$$

letting the vector $\vec{X}$ represent the spatial string coordinates. We then find

$$\begin{aligned} \frac{\partial X^\mu}{\partial \sigma} &= \Big(\frac{\partial X^0}{\partial \sigma}, \frac{\partial \vec{X}}{\partial \sigma}\Big) = \Big(0, \frac{\partial \vec{X}}{\partial \sigma}\Big), \\ \frac{\partial X^\mu}{\partial \tau} &= \Big(\frac{\partial X^0}{\partial t}, \frac{\partial \vec{X}}{\partial t}\Big) = \Big(c\,, \frac{\partial \vec{X}}{\partial t}\Big)\,. \end{aligned} \tag{6.65}$$

As you can see, this parameterization separates the time and space components quite neatly.

Now that we have made a choice of τ coordinates, we can do a simple test to confirm that we got the right sign under the radical in the Nambu–Goto action (6.39). Imagine a little piece of string with no velocity. Because it is not moving, $\partial \vec{X}/\partial t = 0$, and using (6.65), the square root in (6.39) becomes

$$\sqrt{0 - \left(\frac{\partial \vec{X}}{\partial \sigma}\right)^2 (-c^2)}. \tag{6.66}$$

The quantity under the square root is positive, just as we expected. If some day you forget the sign under the radical in the string action, this is a good way to check it quickly.

6.7 Tension and energy of a stretched string

Let us now do our first calculation with the Nambu–Goto action – our first calculation in string theory! We are going to analyze a stretched relativistic string. The endpoints of the string are fixed at $x^1 = 0$, and at $x^1 = a > 0$, with vanishing values for the coordinates of the additional spatial dimensions. We therefore denote the spatial coordinates of the endpoints as $(0, \vec{0})$ and $(a, \vec{0})$. The inclusion of the common $(d-1)$-dimensional vector $\vec{0}$ tells us that the string is only stretched along the first spatial coordinate.

We evaluate the string action for this stretched string using the static gauge $X^0 = c\tau$. Because this is a static string stretched from $x^1 = 0$ to $x^1 = a$, we can write

$$X^1(t, \sigma) = f(\sigma)\,, \quad X^2 = X^3 = \cdots = X^d = 0\,, \tag{6.67}$$

where

$$f(0) = 0\,, \quad f(\sigma_1) = a\,, \tag{6.68}$$

and the function $f(\sigma)$ is strictly increasing and continuous on the interval $\sigma \in [0, \sigma_1]$. The setup is illustrated in Figure 6.9. The function f must be strictly increasing to ensure that each point along the string is assigned a unique σ coordinate.

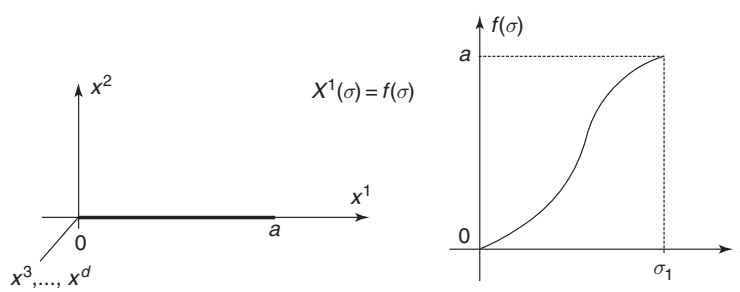

Fig. 6.9 A string of length a stretched along the x^1 axis. The string is parameterized as $X^1(t, \sigma) = f(\sigma)$.

It now follows that

$$\dot{X}^\mu = (c,\, 0,\, \vec{0}), \quad X'^\mu = (0,\, f',\, \vec{0}), \tag{6.69}$$

with $f' = df/d\sigma > 0$. Therefore

$$(\dot{X})^2 = -c^2\,, \quad (X')^2 = (f')^2\,, \quad \dot{X}\cdot X' = 0. \tag{6.70}$$

We can now evaluate the action (6.39):

$$S = -\frac{T_0}{c}\int_{t_i}^{t_f} dt \int_0^{\sigma_1} d\sigma \sqrt{0-(-c^2)(f')^2} = -T_0 \int_{t_i}^{t_f} dt \int_0^{\sigma_1} d\sigma\, \frac{df}{d\sigma}\,. \tag{6.71}$$

The σ integrand is a total derivative, so

$$S = -T_0 \int_{t_i}^{t_f} dt\, (f(\sigma_1) - f(0)) = \int_{t_i}^{t_f} dt\, (-T_0 a)\,, \tag{6.72}$$

where we used (6.68). Note that the value of the action does not depend on the function f used to parameterize the string. This is an explicit confirmation of the reparameterization invariance of the string action.

We would like to interpret our result. For this, recall that the action is the time integral of the Lagrangian L. When the kinetic energy vanishes, $L = -V$, where V is the potential energy. Since our string is static, there is no kinetic energy, so

$$S = \int_{t_i}^{t_f} dt(-V). \tag{6.73}$$

Comparing this with (6.72) we conclude that

$$V = T_0\, a\,. \tag{6.74}$$

The potential energy of our stretched string is just $T_0 a$. What does this mean? If the tension of a static string is T_0, regardless of its length, then $T_0 a$ is the amount of energy that you must spend to create a string of length a. Imagine that you start with an infinitesimal string and you start pulling it. As you do work you are giving energy to the string, in fact, you are creating rest energy, or rest mass. The rest mass μ_0 per unit length is

$$\mu_0 c^2 = \frac{V}{a} = T_0 \;\longrightarrow\; \mu_0 = \frac{T_0}{c^2}. \tag{6.75}$$

The mass (or rest energy) arises only because the string has a tension. Because of this, the relativistic string is sometimes referred to as a massless string. The above calculation supports the identification of T_0 as the string tension. It also confirms that the minus sign in front of the action (6.39) is necessary – otherwise the potential energy of the stretched string would have come out negative.

There is one point that we have glossed over. We assumed in our analysis that the configuration (6.67) satisfies the string equations of motion. If it does not, then the configuration cannot be physically realized. Let us check that the equations of motion are satisfied.

First note that on account of (6.69) neither $\dot{X}^\mu$ nor $X^{\mu\prime}$ has τ dependence. Therefore neither $\mathcal{P}^\tau$ nor $\mathcal{P}^\sigma$ has τ dependence (see (6.49) and (6.50)). This being the case, the equation of motion (6.53) reduces to

$$\frac{\partial \mathcal{P}^\sigma_\mu}{\partial \sigma} = 0\,. \tag{6.76}$$

This requires that $\mathcal{P}^\sigma_\mu$ be σ-independent. We look again at (6.50) and use (6.70) to find

$$\mathcal{P}^\sigma_\mu = -\frac{T_0}{c}\frac{c^2 X'_\mu}{\sqrt{c^2(f')^2}} = -T_0\frac{X'_\mu}{f'}\,. \tag{6.77}$$

This is nonvanishing only for $\mu = 1$, in which case $X'_1 = f'$, so $\mathcal{P}^\sigma$ is indeed σ-independent. Thus the equation of motion is satisfied. Even the boundary conditions are satisfied. As we discussed in Section 6.5, there is no condition to check for string coordinates that satisfy Dirichlet boundary conditions at the endpoints. In our problem this means that there are no extra conditions to be checked for any of the spatial coordinates. For the zeroth coordinate, equation (6.57) requires the free boundary condition $\mathcal{P}^\sigma_0 = 0$. This holds on account of (6.77).

6.8 Action in terms of transverse velocity

We have chosen a partial parameterization of the world-sheet by imposing the condition $X^0 = ct = c\tau$. With this choice, a line of constant τ on the world-sheet corresponds to the string, as seen by our chosen Lorentz observer, at the particular time $t = \tau$.

Can we define some sort of string velocity? Since the components of $\vec{X}(t,\sigma)$ are the string spatial coordinates, the derivative $\partial\vec{X}/\partial t$ seems to be the closest thing we have to a velocity. This velocity, however, depends upon the choice of σ. Its direction, for example, goes along the lines of constant σ. Since σ can be chosen quite arbitrarily, keeping σ constant in taking the derivative is clearly not very physically significant!

Fixing physically the σ parameterization of a string is subtle because the string is an object with no substructure. When comparing a string at two nearby times, it is not possible to say that a point moved from one location to the next. To speak of points on the string we need a σ parameterization, and reparameterization invariance makes it clear to us that this parameterization is not unique. This suggests that longitudinal motion on the string is not physically meaningful.

There is a reparameterization invariant velocity that can be defined on the string. This is, however, a *transverse* velocity. We consider the string motion in *space*, and imagine that each point on the string moves transversely to the string (Figure 6.10). Consider a string at some fixed time t and pick a point p on it. Draw the hyperplane orthogonal to the string at p. At time $t + dt$, with dt infinitesimal, the string has moved, but it will still intersect the plane, this time at a point p'. The transverse velocity is what we get if we presume that the point p moved to p'. No string parameterization is needed to define this velocity.

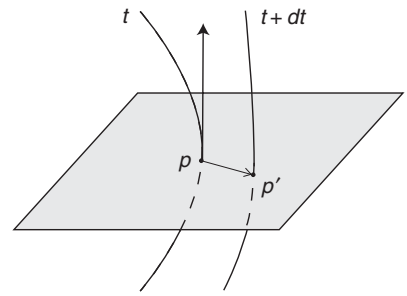

Fig. 6.10 **A string at time t and the plane orthogonal to the string at p. At time $t + dt$ the string intersects the plane at p'. To define the transverse velocity we assume that p moved to p'.**

When speaking of evolving strings there are two surfaces we can discuss. One is the world-sheet, the surface in spacetime which represents the history of the string. The other is a surface in *space*. This *spatial surface* is put together by combining the strings that we observe at all times. This is the surface that would be generated if the string were to leave a wake as it moved. The transverse velocity $\vec{v}_\perp$ at any point on the string is a vector orthogonal to the string and tangent to the string spatial surface. Since $\vec{v}_\perp$ is a string reparameterization invariant notion of velocity, we expect it to enter naturally into the evaluation of the string action.

In order to define the transverse velocity $\vec{v}_\perp$, it is useful to have a unit vector tangent to the string. To this end, we now introduce a parameter s which is more physical than our nearly arbitrary σ: s measures length along the string. Let us work with the string at a *fixed* time, and define $s(\sigma)$ to be the length of the string in the interval $[0, \sigma]$. Thus, for example, $s(0) = 0$, and $s(\sigma_1)$ is the length of an entire open string. Since ds is the length of the infinitesimal vector $d\vec{X}$ which arises from an interval $d\sigma$ along the string, we have:

$$ds = |d\vec{X}| = \left|\frac{\partial \vec{X}}{\partial \sigma}\right| |d\sigma| \,. \tag{6.78}$$

Now consider the quantity $\partial\vec{X}/\partial s$, which is the rate of change of $\vec{X}$ with respect to the length of the string. First note that it is a unit vector:

$$\frac{\partial \vec{X}}{\partial s} \cdot \frac{\partial \vec{X}}{\partial s} = \frac{\partial \vec{X}}{\partial \sigma} \cdot \frac{\partial \vec{X}}{\partial \sigma} \left(\frac{d\sigma}{ds}\right)^2 = \left|\frac{\partial \vec{X}}{\partial \sigma}\right|^2 \left(\frac{d\sigma}{ds}\right)^2 = 1 \,. \tag{6.79}$$

The derivative $\partial\vec{X}/\partial\sigma$ is taken with t held fixed, so it lies along a line of constant t. Since the lines of constant t are precisely the strings, it is tangent to the string. In addition

$$\frac{\partial \vec{X}}{\partial s} = \frac{\partial \vec{X}}{\partial \sigma} \frac{d\sigma}{ds} \,, \tag{6.80}$$

and thus $\partial\vec{X}/\partial s$ is also tangent to the string. Because it has unit length,

$$\frac{\partial \vec{X}}{\partial s} \quad \text{is a unit vector tangent to the string} \,. \tag{6.81}$$

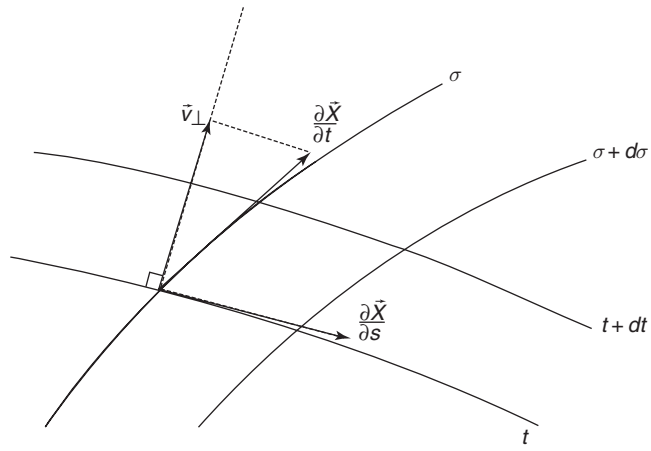

Fig. 6.11 **A small piece of the world-sheet showing the vector $\partial\vec{X}/\partial t$, the transverse velocity $\vec{v}_\perp$, and the unit vector $\partial\vec{X}/\partial s$.**

We define $\vec{v}_\perp$ to be the component of the velocity $\partial\vec{X}/\partial t$, in the direction perpendicular to the string (see Figure 6.11). For any vector $\vec{u}$, its component perpendicular to a unit vector $\vec{n}$ is $\vec{u} - (\vec{u}\cdot\vec{n})\vec{n}$. Therefore, using our unit vector $\partial\vec{X}/\partial s$ along the string, we have

$$\vec{v}_\perp = \frac{\partial\vec{X}}{\partial t} - \Big(\frac{\partial\vec{X}}{\partial t}\cdot\frac{\partial\vec{X}}{\partial s}\Big)\frac{\partial\vec{X}}{\partial s}\,. \tag{6.82}$$

For future use, we calculate $v_\perp^2$. A small computation gives

$$v_\perp^2 = \Big(\frac{\partial\vec{X}}{\partial t}\Big)^2 - \Big(\frac{\partial\vec{X}}{\partial t}\cdot\frac{\partial\vec{X}}{\partial s}\Big)^2\,. \tag{6.83}$$

Our goal now is to write the string action in terms of $\vec{v}_\perp$ and other quantities, if necessary. Using the static gauge $\tau = t$, and equations (6.65), we find

$$(\dot{X})^2 = -c^2 + \Big(\frac{\partial\vec{X}}{\partial t}\Big)^2, \quad (X')^2 = \Big(\frac{\partial\vec{X}}{\partial\sigma}\Big)^2, \quad \dot{X}\cdot X' = \frac{\partial\vec{X}}{\partial t}\cdot\frac{\partial\vec{X}}{\partial\sigma}\,. \tag{6.84}$$

With these relations we simplify the argument of the square root in the string action:

$$\begin{aligned}(\dot{X}\cdot X')^2 - (\dot{X})^2(X')^2 &= \Big(\frac{\partial\vec{X}}{\partial t}\cdot\frac{\partial\vec{X}}{\partial\sigma}\Big)^2 + \Big[c^2 - \Big(\frac{\partial\vec{X}}{\partial t}\Big)^2\Big]\Big(\frac{\partial\vec{X}}{\partial\sigma}\Big)^2 \\ &= \Big(\frac{ds}{d\sigma}\Big)^2\Big[\Big(\frac{\partial\vec{X}}{\partial t}\cdot\frac{\partial\vec{X}}{\partial s}\Big)^2 + c^2 - \Big(\frac{\partial\vec{X}}{\partial t}\Big)^2\Big]. \end{aligned} \tag{6.85}$$

The terms on the right-hand side above can be neatly expressed in terms of $v_\perp^2$. Making use of (6.83),

$$(\dot{X}\cdot X')^2 - (\dot{X})^2(X')^2 = \Big(\frac{ds}{d\sigma}\Big)^2\left(c^2 - v_\perp^2\right), \tag{6.86}$$

or, alternatively,

$$\sqrt{(\dot{X}\cdot X')^2-(\dot{X})^2(X')^2}=c\,\frac{ds}{d\sigma}\sqrt{1-\frac{v_\perp^2}{c^2}}\,. \tag{6.87}$$

This simple expression for the string Lagrangian density shows that $\vec{v}_\perp$ is a natural dynamical variable. Moreover, the longitudinal component of the velocity is completely irrelevant. Now we can write the string action as

$$S=-T_0\int dt\int_0^{\sigma_1}d\sigma\Big(\frac{ds}{d\sigma}\Big)\sqrt{1-\frac{v_\perp^2}{c^2}}\,. \tag{6.88}$$

Here $ds/d\sigma=|\partial\vec{X}/\partial\sigma|$. We did not cancel the $d\sigma$ because it is typically useful to have an integral over a fixed parameter range. While the range of σ is constant, the total length of a string is time dependent. We introduced s as the function of σ that gives length along the string at a fixed time. This definition, used at different times, endows s with a time dependence that can be relevant if we compare strings at different times.

The associated Lagrangian is given by

$$L=-T_0\int ds\sqrt{1-\frac{v_\perp^2}{c^2}}\,. \tag{6.89}$$

This formula was written as an integral over the length parameter in order to give an interpretation. For each infinitesimal piece of string, $T_0 ds$ is its rest energy. As a result, the Lagrangian is an integral over the string of (minus) the rest energy times a local relativistic factor. In this form, we recognize (6.89) as the natural generalization of the relativistic particle Lagrangian (5.8).

The action (6.88) is valid both for open strings and for closed strings. Although relatively simple, it still leads to rather complicated equations of motion in all but the most symmetrical situations. In order to obtain simple equations of motion, we will have to be clever in our choice of σ. For open strings, in addition, we must understand how the endpoints move.

We conclude this section by simplifying our expressions (6.49) and (6.50) for $\mathcal{P}^{\tau\mu}$ and $\mathcal{P}^{\sigma\mu}$ in the static gauge. Let us begin with $\mathcal{P}^{\sigma\mu}$. Its denominator is given in (6.87) and its numerator is simplified using relations (6.84). We find

$$\mathcal{P}^{\sigma\mu}=-\frac{T_0}{c}\,\frac{\left(\dfrac{\partial\vec{X}}{\partial\sigma}\cdot\dfrac{\partial\vec{X}}{\partial t}\right)\dot{X}^\mu-\left(-c^2+\left(\dfrac{\partial\vec{X}}{\partial t}\right)^2\right)X^{\mu\prime}}{c\dfrac{ds}{d\sigma}\sqrt{1-\dfrac{v_\perp^2}{c^2}}}\,. \tag{6.90}$$

Bringing the $ds/d\sigma$ from the denominator up to the numerator, we can turn derivatives with respect to σ into derivatives with respect to s:

$$\mathcal{P}^{\sigma\mu} = -\frac{T_0}{c^2}\frac{\left(\frac{\partial\vec{X}}{\partial s}\cdot\frac{\partial\vec{X}}{\partial t}\right)\dot{X}^\mu + \left(c^2 - \left(\frac{\partial\vec{X}}{\partial t}\right)^2\right)\frac{\partial X^\mu}{\partial s}}{\sqrt{1-\frac{v_\perp^2}{c^2}}}\,. \tag{6.91}$$

The $\mu = 0$ component of this quantity simplifies considerably. Since $\dot{X}^0 = c$ and $\partial X^0/\partial s = c\,\partial t/\partial s = 0$, we find

$$\mathcal{P}^{\sigma 0} = -\frac{T_0}{c}\frac{\left(\frac{\partial\vec{X}}{\partial s}\cdot\frac{\partial\vec{X}}{\partial t}\right)}{\sqrt{1-\frac{v_\perp^2}{c^2}}}\,. \tag{6.92}$$

A rather similar calculation for $\mathcal{P}^{\tau\mu}$ gives

$$\mathcal{P}^{\tau\mu} = \frac{T_0}{c^2}\frac{ds}{d\sigma}\frac{\dot{X}^\mu - \left(\frac{\partial\vec{X}}{\partial s}\cdot\frac{\partial\vec{X}}{\partial t}\right)\frac{\partial X^\mu}{\partial s}}{\sqrt{1-\frac{v_\perp^2}{c^2}}}. \tag{6.93}$$

Quick calculation 6.4 Prove (6.93).

It follows from (6.93) that $\mathcal{P}^{\tau 0}$ and $\vec{\mathcal{P}}^\tau$ are given by

$$\mathcal{P}^{\tau 0} = \frac{T_0}{c}\frac{ds}{d\sigma}\frac{1}{\sqrt{1-\frac{v_\perp^2}{c^2}}}\,, \qquad \vec{\mathcal{P}}^\tau = \frac{T_0}{c^2}\frac{ds}{d\sigma}\frac{\vec{v}_\perp}{\sqrt{1-\frac{v_\perp^2}{c^2}}}\,. \tag{6.94}$$

6.9 Motion of open string endpoints

We will now analyze the motion of the endpoints of an open relativistic string. We consider endpoints that are free to move in all directions. Given our discussion in Section 6.5, this means that we have a space-filling D-brane. Free endpoints are specified by the boundary conditions (6.56), which require the vanishing of $\mathcal{P}^\sigma_\mu$ at the endpoints. We will discover two important properties of the free motion of open string endpoints.

- The endpoints move with the speed of light.
- The endpoints move transversely to the string.

On the interior of the string the notion of a velocity was ambiguous. For the string endpoints, however, the velocity is well defined – there is no ambiguity defining the velocity of

points! Therefore, our statements about endpoint motion have content. In the second statement, motion transverse to the string means that the velocity of an endpoint is orthogonal to the tangent to the string at the endpoint.

To prove the above properties we first recall that $\mathcal{P}^{\sigma 0}$ must vanish at the endpoints. Using (6.92) and noting that the square root in the denominator cannot be infinite, we deduce that the numerator must vanish:

$$\frac{\partial \vec{X}}{\partial s} \cdot \frac{\partial \vec{X}}{\partial t} = 0 \quad \text{at the endpoints} . \tag{6.95}$$

Since $\partial \vec{X}/\partial s$ is a unit vector tangent to the string, and $\partial \vec{X}/\partial t$ is the endpoint velocity, this equation proves that the endpoints move transversely to the string – one of our two claims. In agreement with this interpretation, using (6.95) in (6.82) we see that at the endpoints $\vec{v}_\perp = \partial \vec{X}/\partial t \equiv \vec{v}$. Equation (6.95) actually allows for vanishing endpoint velocity, in which case the transversality property would be trivially satisfied. But this cannot happen; the endpoints move with the speed of light, as we now show.

Using (6.95), we simplify the expression (6.91) for $\mathcal{P}^{\sigma\mu}$ at the endpoints:

$$\mathcal{P}^{\sigma\mu} = -T_0 \sqrt{1 - \frac{v^2}{c^2}} \, \frac{\partial X^\mu}{\partial s} \quad \text{at the endpoints.} \tag{6.96}$$

For the space coordinates, $\mu = 1, \ldots, d$, equation (6.96) gives

$$\vec{\mathcal{P}}^{\sigma} = -T_0 \sqrt{1 - \frac{v^2}{c^2}} \, \frac{\partial \vec{X}}{\partial s} = 0 \quad \text{at the endpoints.} \tag{6.97}$$

Since $\partial \vec{X}/\partial s$ is a unit vector, we conclude that

$$v^2 = c^2. \tag{6.98}$$

Free open string endpoints move with the speed of light.

This conclusion implies that the denominator in (6.92) actually vanishes at the endpoints. Equation (6.95) must still hold, otherwise $\mathcal{P}^{\sigma 0}$ would diverge, rather than vanish at the endpoints. In fact, as we approach the string endpoints, the numerator in (6.92) must vanish faster than the denominator to ensure that the ratio goes to zero.

Problems

Problem 6.1 The induced metric on a two-dimensional surface.

A two-dimensional surface in flat three-dimensional space is described by the height function $z = h(x, y)$, so that points on the surface take the form $(x, y, h(x, y))$. This surface is naturally parameterized by x and y. Recall (6.14), which gives the metric $g_{ij}(\xi)$ induced on a surface parameterized by (ξ^1, ξ^2).

(a) Calculate the components of the metric g_{ij} in terms of h and its derivatives. Write a formula for g_{ij} of the form $g_{ij} = \delta_{ij} + \cdots$.

(b) Write the fully simplified appropriate form of the area integral ($A = \int d\xi^1 d\xi^2 \sqrt{g}$).

Problem 6.2 Metric on S^2 from stereographic parameterization.

Consider a unit sphere S^2 in $\mathbb{R}^3$ centered at the origin: $x^2 + y^2 + z^2 = 1$. Denote a point on the sphere by $\vec{x} = (x, y, z)$. In the stereographic parameterization of the sphere we use parameters ξ^1 and ξ^2 and points on the sphere are (see (6.1)):

$$\vec{x}\,(\xi^1, \xi^2) = \Big(x(\xi^1, \xi^2)\,,\ y(\xi^1, \xi^2)\,,\ z(\xi^1, \xi^2) \Big)\,.$$

Given parameters (ξ^1, ξ^2), the corresponding point on the sphere is that which lies on the line that goes through the north pole $N = (0, 0, 1)$ and the point $(\xi^1, \xi^2, 0)$. Note that the north pole itself is not attained for any finite values of the parameters.

(a) Draw a sketch for the above construction. What are the required ranges for ξ^1 and ξ^2 if we wish to parameterize the full sphere (except for the north pole)?
(b) Calculate the functions $x(\xi^1, \xi^2)$, $y(\xi^1, \xi^2)$, and $z(\xi^1, \xi^2)$.
(c) Calculate the four components of the induced metric $g_{ij}(\xi)$. This is the metric on the sphere, described using the ξ parameters. The algebra is a bit messy, but the result is quite simple (use a symbolic manipulator!).
(d) Check your result by computing the area of the sphere using (6.17).

Problem 6.3 Schwarz inequality in $\mathbb{R}^{1,1}$.

Consider a two-dimensional vector space V with a constant metric such that there is a timelike vector t' ($t'^2 = t' \cdot t' < 0$) and a spacelike vector s' ($s'^2 = s' \cdot s' > 0$).

(a) Show that you can construct vectors t and s such that $t \cdot t = -1$, $s \cdot s = 1$, and $t \cdot s = 0$. [Hint: choose t to be in the direction of t'.] The vectors t and s provide a canonical basis for V, which is now identified as the space $\mathbb{R}^{1,1}$.
(b) Consider now two arbitrary vectors v_1 and v_2 in V. Use the basis in (a) to prove that

$$(v_1 \cdot v_2)^2 \geq v_1^2\, v_2^2\,,$$

where the equality only holds if and only if the vectors v_1 and v_2 are parallel. This result gives some additional perspective on our proof of (6.34).

Problem 6.4 Stretched string and a nonrelativistic limit.

Examine the action (6.88) for a relativistic string with endpoints attached at $(0, \vec{0})$ and $(a, \vec{0})$, as in Section 6.7. Consider the nonrelativistic approximation where $|\vec{v}_\perp| \ll c$ and the oscillations are small (see (4.3)). You may denote by $\vec{y}$ the collection of transverse coordinates $X^2, \ldots, X^d$ and write $\vec{y}(t, x)$, where x is the coordinate corresponding to X^1. Explain why the following relations hold:

$$ds^2 = dx^2 + d\vec{y} \cdot d\vec{y}\,, \quad \vec{v}_\perp \simeq \frac{\partial \vec{y}}{\partial t}\,.$$

Show that the action reduces, up to an additive constant, to the *action* for a nonrelativistic string performing small transverse oscillations. What are the tension and the linear mass density of the resulting string? What is the additive constant?

Problem 6.5 Alternative derivation of the nonrelativistic limit.

Consider the same setup and nonrelativistic approximation discussed in Problem 6.4. This time, however, start your analysis with the Nambu–Goto action (6.39) and work in the static gauge. Moreover, parameterize the strings using $X^1 = x = a\sigma/\sigma_1$. This parameterization is allowed for small oscillations. In fact, it is allowed for any motion in which X^1 is an increasing function along the string.

Problem 6.6 Planar motion for open string with attached endpoints.

Consider the motion of a relativistic open string on the (x, y) plane. The string endpoints are attached to $(x, y) = (0, 0)$ and $(x, y) = (a, 0)$, where $a > 0$. We want to study motions that can be described using the function $y(t, x)$ which gives the vertical displacement of the string for $x \in [0, a]$. Show that the Lagrangian (6.89) can be written in the form:

$$L = -T_0 \int_0^a dx \sqrt{1 + y'^2 - \frac{\dot{y}^2}{c^2}}\,.$$

Here $y' = \partial y/\partial x$ and $\dot{y} = \partial y/\partial t$.

Problem 6.7 Time evolution of a closed circular string.

At $t = 0$, a closed string forms a circle of radius R on the (x, y) plane and has zero velocity. The time development of this string can be studied using the action (6.88). The string will remain circular, but its radius will be a time-dependent function $R(t)$. Give the Lagrangian L as a function of $R(t)$ and its time derivative. Calculate the radius and velocity as functions of time. Sketch the spacetime surface traced by the string in a three-dimensional plot with x, y, and ct axes. [Hint: calculate the Hamiltonian associated with L and use energy conservation.]

Problem 6.8 Covariant analysis of open string endpoint motion.

Use the explicit form of $\mathcal{P}^\sigma_\mu$ to calculate $\mathcal{P}^\sigma_\mu \mathcal{P}^{\sigma\mu}$, and use the result of this calculation to prove that free open string endpoints move with the speed of light.

Problem 6.9 Hamiltonian density for relativistic strings.[†]

Consider the string Lagrangian density $\mathcal{L}$ in the static gauge and written in terms of $\partial_\sigma \vec{X}$ and $\partial_t \vec{X}$. Show that the canonical momentum density $\vec{\mathcal{P}}(t, \sigma)$ is given by

$$\vec{\mathcal{P}}(t, \sigma) \equiv \frac{\partial \mathcal{L}}{\partial(\partial_t \vec{X})} = \frac{T_0}{c^2} \frac{\vec{v}_\perp}{\sqrt{1 - \frac{v_\perp^2}{c^2}}} \frac{ds}{d\sigma}\,.$$

Calculate the Hamiltonian density $\mathcal{H}$, again in terms of $\vec{v}_\perp$ and $\frac{ds}{d\sigma}$. Write the total Hamiltonian as $H = \int d\sigma \mathcal{H} = \int ds(\dots)$ and show that your answer is consistent with the interpretation that the energy of the string arises as the energy of transverse motion of a string whose rest mass arises solely from the tension.

Problem 6.10 Circular string in de Sitter space.

The Nambu–Goto action on a curved spacetime with metric $g_{\mu\nu}(x)$ can be written as in (6.39) with the convention that all dot products use the metric $g_{\mu\nu}$:

$$\dot{X} \cdot X' = g_{\mu\nu}(X)\dot{X}^\mu X'^\nu\,, \quad (\dot{X})^2 = g_{\mu\nu}(X)\dot{X}^\mu \dot{X}^\nu\,, \quad (X')^2 = g_{\mu\nu}(X)X'^\mu X'^\nu\,.$$

Consider strings in an expanding four-dimensional de Sitter spacetime, for which the metric $g_{\mu\nu}$ can be taken to be diagonal, with values

$$g_{00} = -1\,, \quad g_{11} = g_{22} = g_{33} = e^{2Ht}\,.$$

The Hubble constant H has units of one over time so that Ht is dimensionless.

(a) Assume $X^0 \equiv ct = c\tau$ and write $X^\mu = \{X^0, \vec{X}\}$. Write the Nambu–Goto action in terms of t and σ derivatives of $\vec{X}$.

(b) Consider now a circular string on the (x^1, x^2) plane, namely

$$X^1(t,\sigma) = r(t)\cos\sigma\,, \quad X^2(t,\sigma) = r(t)\sin\sigma\,, \quad \sigma \in [0, 2\pi]\,,$$

where $r(t)$ is a radius function to be determined. Use this ansatz to simplify the string action and perform the integration over σ. Write your result as

$$S = \int dt\, L(\dot{r}(t), r(t); t)\,,$$

and determine $L(\dot{r}(t), r(t); t)$, which is explicitly time dependent. Because of the e^{2Ht} factors in the metric, the physically measured radius of the string is actually $R(t) = e^{Ht} r(t)$. Write the Lagrangian in terms of R and $\dot{R}$.

(c) Consider strings with constant R and use the Lagrangian to give the potential $V(R)$ for such strings. Plot this potential and verify that it is well-defined only if $R \leq c/H$. Find a critical point of the potential and the corresponding value of R. Is this static string in stable equilibrium?

(d) Use the Lagrangian for R and $\dot{R}$ to calculate the corresponding Hamiltonian, expressed as a function of R and $\dot{R}$. Simplify your answer. Is this Hamiltonian function conserved for physical motion?

Problem 6.11 Open strings ending on D-branes of various dimensions.

Consider a world with d spatial dimensions. A Dp-brane is an extended object with p spatial dimensions: a p-dimensional hyperplane inside the d-dimensional space. We will examine properties of strings ending on a Dp-brane, where $0 \leq p < d$. The case $p = d$, where the D-brane is space filling was discussed in Section 6.9.

For a Dp-brane, let x^i, with $i = 1, \ldots, p$, correspond to directions on the Dp-brane, and x^a with $a = p+1, \ldots, d$, correspond to directions orthogonal to the Dp-brane. The Dp-brane position is specified by $x^a = 0$, for $a = p+1, \ldots, d$. Open string endpoints must lie on the Dp-brane. Focusing on the $\sigma = 0$ endpoint, we thus have

$$X^a(t, \sigma = 0) = 0\,, \quad a = p+1, \ldots, d\,.$$

The motion of the endpoint along the D-brane directions x^i is free. We work in the static gauge.

(a) State the conditions satisfied by $\mathcal{P}_0^\sigma$, $\mathcal{P}_i^\sigma$, and $\mathcal{P}_a^\sigma$ at the endpoint (no condition is a possibility!).

Prove the following.

(b) All boundary conditions are automatically satisfied if the string ends on a D0-brane.
(c) For a string ending on a D1-brane, the tangent to the string at the endpoint is orthogonal to the D1-brane, and the endpoint velocity is unconstrained.
(d) For a string ending on a Dp-brane, with $p \geq 2$, there are two possibilities:
    (i) the string is orthogonal to the Dp-brane at the endpoint, and the endpoint velocity is unconstrained, or,
    (ii) the string is not orthogonal to the Dp-brane at the endpoint, and the endpoint moves with the speed of light transversely to the string.

7 String parameterization and classical motion

We construct lines of constant σ that are perpendicular to the lines of constant τ and use the energy density carried by the string to fix the σ parameterization completely. The resulting system of string equations of motion includes wave equations and two nonlinear constraints. We find the general solution which describes the motion of open strings with free endpoints. We examine the free evolution of closed strings and find that cusps that appear and later disappear are generic. We discuss cosmic strings and the gravitational lensing they can produce.

7.1 Choosing a σ parameterization

We have already learned a few facts about the motion of relativistic strings. In particular, we learned that free open string endpoints move with the speed of light transversally to the string. This result was obtained using the static gauge $X^0 \equiv ct = c\tau$, which partially fixed the parameterization of the world-sheet. Once we have chosen this gauge, the string motion is defined by the functions $\vec{X}(t, \sigma)$. As we vary t and σ, $\vec{X}(t, \sigma)$ describes the string spatial surface – the surface in space consisting of the strings at all times. In this chapter, the string spatial surface will simply be called the string surface. We will also use the static gauge throughout. The statement that the open string endpoints move transversely to the string implies that the vectors tangent to the boundary of the string surface are orthogonal to the strings. Our goal is to find a useful σ parameterization of the string surface. If we know this parameterization, we also know the σ parameterization of the world-sheet and, in fact, the complete parameterization of the world-sheet.

We will now show how to use a particular σ parameterization of a single string to construct a useful σ parameterization of all strings, and thus of the entire string surface. Suppose that the $t = 0$ string is given some σ parameterization with $\sigma \in [0, \sigma_1]$ (see Figure 7.1). Now, consider the string at $t = \epsilon$, with ϵ infinitesimal. On the string surface we can draw short segments perpendicular to the $t = 0$ string. Let these segments intersect the $t = \epsilon$ string. Consider a point σ_0 on the $t = 0$ string, and the short perpendicular above it. We declare the intersection of this perpendicular with the $t = \epsilon$ string to also have $\sigma = \sigma_0$. We do this all over the $t = 0$ string, obtaining a parameterization of the $t = \epsilon$ string. We then repeat this procedure, using the $t = \epsilon$ string to parameterize the $t = 2\epsilon$ string. We continue in this way, working in the limit of very small ϵ. The result is a set of lines of constant σ that are

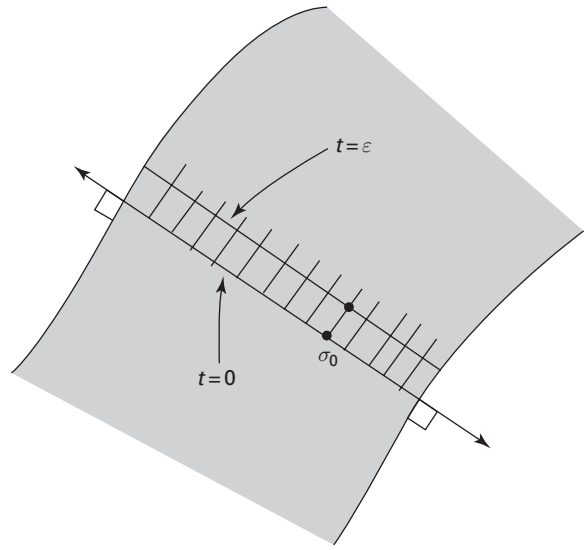

Fig. 7.1 **Using the parameterization of the $t = 0$ string to construct the parameterization of the $t = \epsilon$ string. On the string surface, the lines of constant σ are chosen to be orthogonal to the lines of constant t.**

everywhere orthogonal to the strings (that is, orthogonal to the lines of constant t). This construction can be done both for open and for closed strings. For closed strings the σ range $[0, \sigma_c]$ of the $t = 0$ string becomes the range of all other strings. For open strings the σ range $[0, \sigma_1]$ of the $t = 0$ string also becomes the range of all other strings. This happens because the boundaries of the string surface are orthogonal to the strings, and, as a result, the boundaries are lines of constant σ.

In summary, the σ parameterization of a given string can be used to construct lines of constant σ that are always perpendicular to the lines of constant t. In this parameterization of the string surface, the tangent $\partial\vec{X}/\partial\sigma$ to the strings and the tangent $\partial\vec{X}/\partial t$ to the lines of constant σ are perpendicular to each other at any point:

$$\frac{\partial\vec{X}}{\partial\sigma}\cdot\frac{\partial\vec{X}}{\partial t} = 0. \tag{7.1}$$

Since the velocity $\partial\vec{X}/\partial t$ is perpendicular to the string, it coincides with $\vec{v}_\perp$ (see (6.82)):

$$\vec{v}_\perp = \frac{\partial\vec{X}}{\partial t} \quad \text{at all points,} \tag{7.2}$$

and not only at the endpoints, where it happens independently of parameterization. Recalling that s is the length parameter along the string, equation (7.1) implies that $\frac{\partial \vec{X}}{\partial s} \cdot \frac{\partial \vec{X}}{\partial t} = 0$, which allows us to simplify our earlier results for $\mathcal{P}^{\tau\mu}$ and $\mathcal{P}^{\sigma\mu}$. We find that (6.93) becomes

$$\mathcal{P}^{\tau\mu} = \frac{T_0}{c^2} \frac{\frac{ds}{d\sigma}}{\sqrt{1 - \frac{v_\perp^2}{c^2}}} \frac{\partial X^\mu}{\partial t}. \tag{7.3}$$

Similarly, (6.91) becomes

$$\mathcal{P}^{\sigma\mu} = -T_0 \sqrt{1 - \frac{v_\perp^2}{c^2}}\, \frac{\partial X^\mu}{\partial s}. \tag{7.4}$$

Equation (7.4) holds at the open string endpoints regardless of σ-parameterization (see (6.96)); with the present parameterization it holds all over the string.

7.2 Physical interpretation of the string equation of motion

Having used the σ parameterization to simplify some of our previous expressions, we now look into the string equations of motion (6.53). With $t = \tau$ we have

$$\frac{\partial \mathcal{P}^{\tau\mu}}{\partial t} = -\frac{\partial \mathcal{P}^{\sigma\mu}}{\partial \sigma}. \tag{7.5}$$

Let us first consider the $\mu = 0$ component of this equation. It follows from (7.4) that $\mathcal{P}^{\sigma 0} = 0$. Furthermore, equation (7.3) gives

$$\mathcal{P}^{\tau 0} = \frac{T_0}{c} \frac{\frac{ds}{d\sigma}}{\sqrt{1 - \frac{v_\perp^2}{c^2}}}. \tag{7.6}$$

Back into the equation of motion (7.5), we obtain

$$\frac{\partial \mathcal{P}^{\tau 0}}{\partial t} = \frac{\partial}{\partial t} \left(\frac{T_0 \frac{ds}{d\sigma}}{c\sqrt{1 - \frac{v_\perp^2}{c^2}}} \right) = 0. \tag{7.7}$$

To understand this result physically, consider a little piece of string associated with a $d\sigma$ that is a small fixed number. The motion of this particular piece of string is well defined now that we have fixed the lines of constant σ. Since ds denotes the length of the $d\sigma$ piece of string, ds can depend on time. If we multiply equation (7.7) by the constant $d\sigma$, we conclude that

$$\frac{T_0\, ds}{\sqrt{1 - \frac{v_\perp^2}{c^2}}}, \tag{7.8}$$

is constant in time. Expression (7.8) has units of energy, which suggests that it may be the relativistic energy associated with this piece of the string. Indeed, this is in agreement with our findings in Section 6.7 where we saw that the rest energy of a static stretched string is given by its length times the tension T_0. The rest energy in the above expression is $T_0 ds$, and the relativistic factor in the denominator makes (7.8) the total energy. Equation (7.7) therefore states that the energy in *each* piece $d\sigma$ of the string is conserved. This is a very interesting fact. It implies, for example, that the energy in the range $[0, \sigma_0]$ of an open string is constant in time for any fixed σ_0.

The above interpretation is also confirmed by a calculation of the energy E of a relativistic string (Problem 6.9). You found that the Hamiltonian is

$$H = \int \frac{T_0\, ds}{\sqrt{1 - \frac{v_\perp^2}{c^2}}}, \tag{7.9}$$

which shows that (7.8) is the energy in a piece of string.

Now we turn to the space components of the string equation of motion. The space components $\vec{\mathcal{P}}^\tau$ of $\mathcal{P}^{\tau\mu}$ can be read from (7.3):

$$\vec{\mathcal{P}}^\tau = \frac{T_0}{c^2} \frac{\frac{ds}{d\sigma}}{\sqrt{1 - \frac{v_\perp^2}{c^2}}} \vec{v}_\perp, \tag{7.10}$$

and similarly, from (7.4),

$$\vec{\mathcal{P}}^\sigma = -T_0 \sqrt{1 - \frac{v_\perp^2}{c^2}}\, \frac{\partial \vec{X}}{\partial s}. \tag{7.11}$$

Now we can substitute these expressions back into (7.5) to find that

$$\begin{aligned} \frac{\partial}{\partial \sigma}\left[T_0 \sqrt{1 - \frac{v_\perp^2}{c^2}}\, \frac{\partial \vec{X}}{\partial s}\right] &= \frac{\partial}{\partial t}\left[\frac{T_0}{c^2} \frac{\frac{ds}{d\sigma}}{\sqrt{1 - \frac{v_\perp^2}{c^2}}} \vec{v}_\perp\right] \\ &= \frac{T_0}{c^2} \frac{\frac{ds}{d\sigma}}{\sqrt{1 - \frac{v_\perp^2}{c^2}}} \frac{\partial \vec{v}_\perp}{\partial t}, \end{aligned} \tag{7.12}$$

where the final step used equation (7.7). It is possible to interpret this equation loosely in terms of an "effective" nonrelativistic string. Recall that the equations of motion for a classical nonrelativistic string are

$$\mu_0 \frac{\partial^2 \vec{y}}{\partial t^2} = T_0 \frac{\partial^2 \vec{y}}{\partial x^2} = \frac{\partial}{\partial x}\left[T_0 \frac{\partial \vec{y}}{\partial x}\right], \tag{7.13}$$

where x is a length parameter along the direction defined by the static stretched string, and $\vec{y}$ is the transverse displacement. How do we recast (7.12) to resemble (7.13)? We use the

$ds/d\sigma$ factor on the right-hand side to transform the σ-derivative on the left-hand side into an s-derivative:

$$\frac{T_0}{c^2}\frac{1}{\sqrt{1-\frac{v_\perp^2}{c^2}}}\frac{\partial \vec{v}_\perp}{\partial t} = \frac{\partial}{\partial s}\left[T_0\sqrt{1-\frac{v_\perp^2}{c^2}}\left(\frac{\partial \vec{X}}{\partial s}\right)\right]. \tag{7.14}$$

For small oscillations the length parameter s along the vibrating string is roughly equal to the parameter x along the direction of the static string. We can then compare this equation with (7.13) and conclude that the relativistic string has a velocity-dependent effective tension T_{eff} and a velocity-dependent effective mass density μ_{eff} given by

$$T_{\text{eff}} = T_0\sqrt{1-\frac{v_\perp^2}{c^2}}, \quad \mu_{\text{eff}} = \frac{T_0}{c^2}\frac{1}{\sqrt{1-\frac{v_\perp^2}{c^2}}}. \tag{7.15}$$

Since free open string endpoints move with $v_\perp = c$, the effective tension of the string goes to zero at the endpoints. One could say that the endpoints have to move at the speed of light in order to make the tension vanish at the endpoints. This is the only way that the relativistic string can have a tension and still make sense with free open ends. The effective mass density diverges at the endpoints. This is not a problem; the same divergence is present for the energy density which appears as the integrand in (7.9). Despite the singular behavior at the endpoints, the integral turns out to be finite, as required by consistency since, after all, we are describing strings with finite energy.

7.3 Wave equation and constraints

Equation (7.14) is still fairly complicated. It may seem that, having fixed the lines of constant σ, we have run out of reparameterizations that can help us simplify the equations of motion. This is not the case, however. We showed how to construct the lines of constant σ *if* we have one parameterized string. We must now try to parameterize this first string in the best possible way!

Here is the physical way to do so: we will parameterize the string so that each string segment of equal parameter length σ carries the same amount of energy. We will parameterize the string using the energy! This parameterization will yield simple equations of motion. To see this, we first rewrite (7.14) suggestively by changing s-derivatives into σ-derivatives:

$$\frac{1}{c^2}\frac{\partial^2 \vec{X}}{\partial t^2} = \frac{\sqrt{1-\frac{v_\perp^2}{c^2}}}{\frac{ds}{d\sigma}}\frac{\partial}{\partial \sigma}\left[\frac{\sqrt{1-\frac{v_\perp^2}{c^2}}}{\frac{ds}{d\sigma}}\frac{\partial \vec{X}}{\partial \sigma}\right]. \tag{7.16}$$

Let $A(\sigma)$ denote the ratio that appears prominently in the above equation:

$$A(\sigma) = \frac{\frac{ds}{d\sigma}}{\sqrt{1 - \frac{v_\perp^2}{c^2}}}. \tag{7.17}$$

We already showed in (7.7) that $A(\sigma)$ is independent of time. We will now choose σ in such a way that $A = 1$. If we do so, the equation of motion (7.16) becomes the familiar wave equation:

$$\frac{1}{c^2}\frac{\partial^2 \vec{X}}{\partial t^2} = \frac{\partial^2 \vec{X}}{\partial \sigma^2}. \tag{7.18}$$

This is a great simplification. To find a σ that leads to $A = 1$ we assign $\sigma = 0$ to one endpoint of the open string and work our way along the string assigning to each piece ds of string the interval $d\sigma$ given by

$$d\sigma = \frac{ds}{\sqrt{1 - \frac{v_\perp^2}{c^2}}} = \frac{1}{T_0} dE. \tag{7.19}$$

The first equality implies $A = 1$, and the second equality follows from the identification of (7.8) with the energy dE carried by the little piece of string. In this parameterization the energy density $dE/d\sigma$ is a constant equal to the tension. Equation (7.19) can be integrated from the $\sigma = 0$ endpoint up to a point Q, giving

$$\sigma(Q) = \frac{E(Q)}{T_0}. \tag{7.20}$$

The coordinate $\sigma(Q)$ assigned to Q equals the energy $E(Q)$ carried by the portion of the string stretching from the selected endpoint up to Q, divided by the tension. It also follows from the above equation that

$$\sigma \in [0, \sigma_1\,], \quad \sigma_1 = \frac{E}{T_0}, \tag{7.21}$$

where E is the total energy of the string.

Our choice of σ, done for all strings, is consistent with the orthogonality condition (7.1). In fact, the lines of constant σ have not been changed, only the values of σ assigned to them have. Our constant σ lines still ensure that the energy in the $[0, \sigma\,]$ portion of the strings is constant and such lines are orthogonal to the strings.

The parameterization condition (7.19) is actually equivalent to a differential constraint on the coordinates $\vec{X}$. We first rewrite the first equality in (7.19) as

$$\left(\frac{ds}{d\sigma}\right)^2 + \frac{1}{c^2} v_\perp^2 = 1. \tag{7.22}$$

Recalling that $\partial \vec{X}/\partial s$ is a unit vector, and making use of (7.2), we find

$$\left(\frac{\partial \vec{X}}{\partial \sigma}\right)^2 + \frac{1}{c^2}\left(\frac{\partial \vec{X}}{\partial t}\right)^2 = 1. \tag{7.23}$$

Finally, let us examine the boundary conditions. From (7.11) we have

$$\vec{\mathcal{P}}^{\sigma} = -T_0\sqrt{1-\frac{v_\perp^2}{c^2}}\,\frac{d\sigma}{ds}\,\frac{\partial\vec{X}}{\partial\sigma} = -T_0\,\frac{\partial\vec{X}}{\partial\sigma}. \tag{7.24}$$

Therefore, the free endpoint boundary condition is very simple:

$$\frac{\partial\vec{X}}{\partial\sigma} = 0 \quad \text{at the endpoints.} \tag{7.25}$$

With the chosen parameterization, the free endpoint boundary condition, first introduced in (6.56), is simply a Neumann boundary condition.

All in all we have four equations to solve in order to find the motion of a relativistic string: (7.18), (7.1), (7.23), and (7.25). We collect them here:

$$\text{wave equation:} \quad \frac{\partial^2\vec{X}}{\partial\sigma^2} - \frac{1}{c^2}\frac{\partial^2\vec{X}}{\partial t^2} = 0, \tag{7.26}$$

$$\text{parameterization condition:} \quad \frac{\partial\vec{X}}{\partial t}\cdot\frac{\partial\vec{X}}{\partial\sigma} = 0\,, \tag{7.27}$$

$$\text{parameterization condition:} \quad \Big(\frac{\partial\vec{X}}{\partial\sigma}\Big)^2 + \frac{1}{c^2}\Big(\frac{\partial\vec{X}}{\partial t}\Big)^2 = 1, \tag{7.28}$$

$$\text{boundary condition:} \quad \frac{\partial\vec{X}}{\partial\sigma}\Big|_{\sigma=0} = \frac{\partial\vec{X}}{\partial\sigma}\Big|_{\sigma=\sigma_1} = 0. \tag{7.29}$$

For a string with energy E, the above equations require $\sigma_1 = E/T_0$. For the record, we also include (7.19), written as

$$\frac{1}{T_0}\frac{dE}{d\sigma} = \frac{\frac{ds}{d\sigma}}{\sqrt{1-\frac{v_\perp^2}{c^2}}} = 1. \tag{7.30}$$

Finally, from equations (7.3) and (7.4) we get

$$\mathcal{P}^{\tau\mu} = \frac{T_0}{c^2}\,\frac{\partial X^\mu}{\partial t}, \tag{7.31}$$

$$\mathcal{P}^{\sigma\mu} = -T_0\,\frac{\partial X^\mu}{\partial\sigma}. \tag{7.32}$$

7.4 General motion of an open string

In this section our goal is to describe the general motion of open strings with free boundary conditions. We will therefore examine in detail how to solve equations (7.26)–(7.29).

Let us first consider the wave equation for $\vec{X}$. This equation is readily solved in terms of arbitrary vector functions of $(ct \pm \sigma)$. We thus write

$$\vec{X}(t,\sigma) = \frac{1}{2}\big(\vec{F}(ct+\sigma) + \vec{G}(ct-\sigma)\big). \tag{7.33}$$

The boundary condition at the $\sigma = 0$ endpoint demands that

$$\frac{\partial \vec{X}}{\partial \sigma}\Big|_{\sigma=0} = 0 \longrightarrow \vec{F}'(ct) - \vec{G}'(ct) = 0, \tag{7.34}$$

where prime denotes derivative with respect to the argument. Since ct takes all possible values, the above equation holds for all values of the argument. Calling u the argument, we write

$$\frac{d\vec{F}(u)}{du} = \frac{d\vec{G}(u)}{du} \longrightarrow \vec{G}(u) = \vec{F}(u) + \vec{a}_0, \tag{7.35}$$

where $\vec{a}_0$ is a constant vector. Back in (7.33) we now have

$$\vec{X}(t,\sigma) = \frac{1}{2}\Big(\vec{F}(ct+\sigma) + \vec{F}(ct-\sigma) + \vec{a}_0\Big). \tag{7.36}$$

We can absorb the constant vector $\vec{a}_0$ into the definition of $\vec{F}$ (call $\vec{F}(u) + \vec{a}_0/2$ the new $\vec{F}$), and therefore our solution so far takes the form

$$\vec{X}(t,\sigma) = \frac{1}{2}\Big(\vec{F}(ct+\sigma) + \vec{F}(ct-\sigma)\Big). \tag{7.37}$$

Consider now the boundary condition at $\sigma = \sigma_1$:

$$\frac{\partial \vec{X}}{\partial \sigma}\Big|_{\sigma=\sigma_1} = 0 \longrightarrow \vec{F}'(ct+\sigma_1) - \vec{F}'(ct-\sigma_1) = 0. \tag{7.38}$$

Letting $u = ct - \sigma_1$, the above condition becomes

$$\frac{d\vec{F}}{du}(u + 2\sigma_1) = \frac{d\vec{F}}{du}(u). \tag{7.39}$$

This equation tells us that the derivative of $\vec{F}$ is periodic with period $2\sigma_1$. Integrating, we find that the function $\vec{F}$ is quasi-periodic: after a period $2\sigma_1$ it changes by a fixed constant. We write this as

$$\vec{F}(u + 2\sigma_1) = \vec{F}(u) + 2\sigma_1 \frac{\vec{v}_0}{c}, \tag{7.40}$$

where $\vec{v}_0$ is a vector constant of integration with units of velocity, and the constants have been added for convenience. This concludes our analysis of the boundary conditions.

We now examine what restrictions the parameterization conditions (7.27) and (7.28) impose on the function $\vec{F}$. A standard trick is to add and subtract the first equation to the second, as in

$$\Big(\frac{\partial \vec{X}}{\partial \sigma}\Big)^2 \pm 2\,\frac{\partial \vec{X}}{\partial \sigma}\cdot\frac{1}{c}\frac{\partial \vec{X}}{\partial t} + \frac{1}{c^2}\Big(\frac{\partial \vec{X}}{\partial t}\Big)^2 = 1, \tag{7.41}$$

which can then be written more briefly as

$$\left(\frac{\partial \vec{X}}{\partial \sigma} \pm \frac{1}{c}\frac{\partial \vec{X}}{\partial t}\right)^2 = 1. \tag{7.42}$$

Note that this is equivalent to the two constraints (7.27) and (7.28). Using (7.37) we can evaluate the derivatives that enter into the constraints:

$$\begin{aligned} \frac{\partial \vec{X}}{\partial \sigma} &= \frac{1}{2}\big(\vec{F}'(ct+\sigma) - \vec{F}'(ct-\sigma)\big), \\ \frac{1}{c}\frac{\partial \vec{X}}{\partial t} &= \frac{1}{2}\big(\vec{F}'(ct+\sigma) + \vec{F}'(ct-\sigma)\big). \end{aligned} \tag{7.43}$$

As a result,

$$\frac{\partial \vec{X}}{\partial \sigma} \pm \frac{1}{c}\frac{\partial \vec{X}}{\partial t} = \pm\vec{F}'(ct \pm \sigma). \tag{7.44}$$

It follows that the constraints (7.42) require $\vec{F}' \cdot \vec{F}' = 1$ for all arguments of $\vec{F}$. In other words, the vector $\vec{F}'(u)$ is a unit vector:

$$\left|\frac{d\vec{F}(u)}{du}\right|^2 = 1. \tag{7.45}$$

This is good progress: the constraints give a simple condition for $\vec{F}(u)$. Since $\vec{F}$ is a vector function of the parameter u, we can visualize $\vec{F}(u)$ as a parameterized curve in space. Equation (7.45) has a simple interpretation:

$$u \text{ is a length parameter along the curve } \vec{F}(u). \tag{7.46}$$

This is explained as follows. Consider two nearby points $\vec{F}(u+du)$ and $\vec{F}(u)$ on the curve. Their vector separation $d\vec{F} = \vec{F}(u+du) - \vec{F}(u)$ has length $|d\vec{F}|$. Equation (7.45) implies that $|d\vec{F}| = |du|$, showing that the parameter change $|du|$ is the distance between the two nearby points.

We can now summarize our analysis of the equations of motion. The general solution which describes the motion of an open string with free endpoints is given by

$$\vec{X}(t,\sigma) = \frac{1}{2}\Big(\vec{F}(ct+\sigma) + \vec{F}(ct-\sigma)\Big), \quad \sigma \in [0,\sigma_1], \tag{7.47}$$

where $\sigma_1 = E/T_0$, E is the energy of the string, and $\vec{F}$ satisfies the conditions

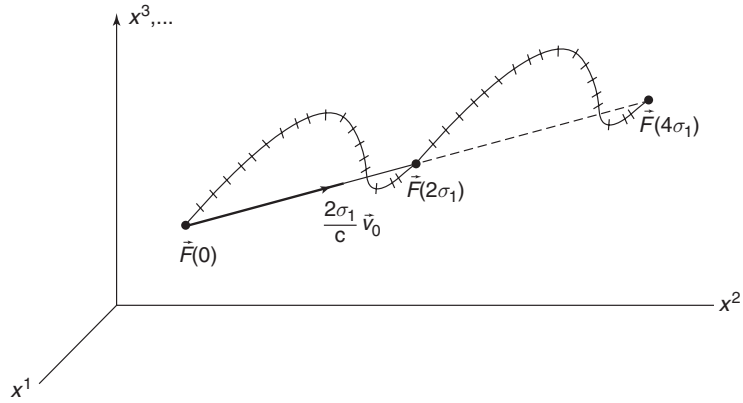

Fig. 7.2 The vector function $\vec{F}(u)$ changes by a constant vector $(2\sigma_1\vec{v}_0/c)$ when $u \to u + 2\sigma_1$. The parameterized curve $\vec{F}(u)$ encodes the full motion of an open string with free endpoints.

$$\left|\frac{d\vec{F}(u)}{du}\right|^2 = 1 \quad \text{and} \quad \vec{F}(u + 2\sigma_1) = \vec{F}(u) + 2\sigma_1\frac{\vec{v}_0}{c}. \tag{7.48}$$

The problem has been reduced to that of finding a vector function $\vec{F}$ which satisfies equations (7.48). The second of these equations tells us that it suffices to find $\vec{F}(u)$ for $u \in [0, 2\sigma_1]$. This determines $\vec{F}(u)$ for all u, and thus it determines $\vec{X}(t, \sigma)$ completely. The interpretation of $\vec{v}_0$ will be given below. An illustration of $\vec{F}$ is shown in Figure 7.2.

We can give a physical interpretation to $\vec{F}(u)$. It follows from (7.47) that the motion of the $\sigma = 0$ endpoint of the open string is described by

$$\vec{X}(t, 0) = \vec{F}(ct). \tag{7.49}$$

Therefore, we see that:

$$\vec{F}(u) \text{ is the position of the } \sigma = 0 \text{ endpoint at time } u/c. \tag{7.50}$$

Additionally, we can give a physical interpretation to the constant velocity $\vec{v}_0$. From the second equation in (7.48) we have

$$\vec{F}(2\sigma_1) = \vec{F}(0) + 2\sigma_1\frac{\vec{v}_0}{c}, \tag{7.51}$$

and expressing the $\vec{F}$ in terms of the position of the $\sigma = 0$ endpoint, we find

$$\vec{X}\left(t = \frac{2\sigma_1}{c}, 0\right) = \vec{X}(t = 0, 0) + \left(\frac{2\sigma_1}{c}\right)\vec{v}_0. \tag{7.52}$$

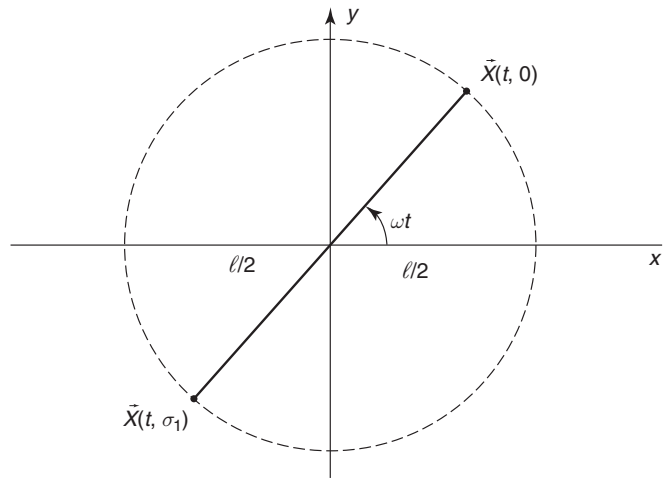

Fig. 7.3 An open string of length ℓ rotating in the (x, y) plane with angular velocity ω.

This shows that $\vec{v}_0$ is the average velocity of the $\sigma = 0$ endpoint during the time interval $[0, 2\sigma_1/c]$.

Quick calculation 7.1 Show that $\vec{v}_0$ is, in fact, the average velocity of any point σ on the string calculated over any time interval of duration $2\sigma_1/c$.

Quick calculation 7.2 Show that the velocity of any point on the string is a periodic function of time with period $2\sigma_1/c$, namely, $\dot{\vec{X}}(t, \sigma) = \dot{\vec{X}}(t + \frac{2\sigma_1}{c}, \sigma)$.

Since $\vec{F}$ can be reconstructed by looking only at the $\sigma = 0$ endpoint, we may ask: how long do we need to observe this endpoint in order to determine the full motion, past and future, of an open string with energy E? Since the motion is determined if we know $\vec{F}(u)$ from $u = 0$ to $u = 2\sigma_1$, we must observe $\vec{X}(t, 0)$ from $t = 0$ to $t = 2\sigma_1/c$. Since $\sigma_1 = E/T_0$, we need to observe the endpoint for a time interval $\Delta t = 2E/cT_0$. This is twice the time that light takes to travel a length E/T_0, the length of a static string of energy E.

We now use the above construction to describe the motion of a straight open string with energy E which rotates rigidly about its fixed midpoint in the (x, y) plane (Figure 7.3). Our first goal is to produce the function $\vec{F}(u)$. This function is easily constructed because we know all about the motion of the endpoints. Assuming that the string is of length ℓ and rotates with angular frequency ω, we describe the motion of the $\sigma = 0$ endpoint by

$$\vec{X}(t, 0) = \frac{\ell}{2}\big(\cos\omega t, \sin\omega t\big), \tag{7.53}$$

where we use vector notation with two components, since the motion is restricted to the (x, y) plane. Given that $\vec{F}(ct) = \vec{X}(t, 0)$ we have

$$\vec{F}(u) = \frac{\ell}{2}\Big(\cos\frac{\omega u}{c}, \sin\frac{\omega u}{c}\Big). \tag{7.54}$$

The function $\vec{F}$ is periodic, suggesting that the vector $\vec{v}_0$ in (7.48) must vanish. More precisely, $\vec{v}_0$ vanishes because it is the average velocity during the period in which the velocity repeats, the velocity repeats after precisely one full turn, and the average velocity of any point during one full turn is zero. With $\vec{v}_0 = 0$ equation (7.48) imposes the condition $\vec{F}(u + 2\sigma_1) = \vec{F}(u)$. Using (7.54), this gives

$$\frac{\omega}{c}(2\sigma_1) = 2\pi m \longrightarrow \frac{\omega}{c} = \frac{\pi}{\sigma_1} m, \tag{7.55}$$

where m is an integer. It is simple to see that we must choose $m = 1$. For this we just calculate $\vec{X}(0, \sigma)$, which gives the string at time equal to zero:

$$\vec{X}(0, \sigma) = \frac{1}{2}\left(\vec{F}(\sigma) + \vec{F}(-\sigma)\right) = \frac{\ell}{2}\Big(\cos\frac{\pi m\sigma}{\sigma_1}, 0\Big). \tag{7.56}$$

For $m = 1$, the string is recovered when $\sigma \in [0, \sigma_1]$. For arbitrary m, the function $\vec{X}(0, \sigma)$ traces out the string m times when $\sigma \in [0, \sigma_1]$. Choosing $m = 1$, we find

$$\frac{\omega}{c} = \frac{\pi}{\sigma_1} = \frac{\pi T_0}{E}. \tag{7.57}$$

This gives the angular frequency of the motion in terms of the energy. The first condition in (7.48) determines the length ℓ. Indeed,

$$\frac{d\vec{F}}{du} = \frac{\omega\ell}{2c}\Big(-\sin\frac{\omega u}{c}, \cos\frac{\omega u}{c}\Big), \tag{7.58}$$

and, as a result,

$$\Big|\frac{d\vec{F}}{du}\Big|^2 = \Big(\frac{\omega\ell}{2c}\Big)^2 = 1 \longrightarrow \ell = \frac{2c}{\omega} = \frac{2\sigma_1}{\pi} = \frac{2}{\pi}\frac{E}{T_0}. \tag{7.59}$$

This length is smaller, by a factor of $2/\pi$, than the length of a static string with energy E. This is sensible since this string has kinetic energy. Alternatively,

$$E = \frac{\pi}{2} T_0 \ell, \tag{7.60}$$

which states that the energy of a rotating string is a factor of $\pi/2$ bigger than the energy of a static string with the same length. Note also that $\omega(\ell/2) = c$, telling us that the endpoints move with the speed of light. In this solution the energy is proportional to the length of the string. Perhaps surprisingly, as the energy of the string increases, the angular velocity ω decreases. This happens because the string endpoints have to move at the speed of light, so as the length of the string grows the angular velocity has to decrease.

Having determined ω and ℓ, we really know the motion of the string. It is of interest, however, to write the complete expression for the motion of the *parameterized* string $\vec{X}(t, \sigma)$. In terms of σ_1, the vector $\vec{F}$ in (7.54) is now given by

$$\vec{F}(u) = \frac{\sigma_1}{\pi}\Big(\cos\frac{\pi u}{\sigma_1}, \sin\frac{\pi u}{\sigma_1}\Big). \tag{7.61}$$

Finally, using (7.47) we obtain

$$\vec{X}(t,\sigma) = \frac{\sigma_1}{2\pi}\left(\cos\frac{\pi(ct+\sigma)}{\sigma_1} + \cos\frac{\pi(ct-\sigma)}{\sigma_1}, \sin\frac{\pi(ct+\sigma)}{\sigma_1} + \sin\frac{\pi(ct-\sigma)}{\sigma_1}\right), \tag{7.62}$$

and after simplification,

$$\vec{X}(t,\sigma) = \frac{\sigma_1}{\pi}\cos\frac{\pi\sigma}{\sigma_1}\left(\cos\frac{\pi ct}{\sigma_1},\ \sin\frac{\pi ct}{\sigma_1}\right). \tag{7.63}$$

The parameterized string is of interest because the energy density is a constant function of the parameter σ. In Problem 7.2 you will calculate the energy $\mathcal{E}(s)$ per unit length on the string as a function of the distance s to the center:

$$\mathcal{E}(s) = \frac{T_0}{\sqrt{1-\frac{4s^2}{\ell^2}}}. \tag{7.64}$$

At the center of the string the energy density $\mathcal{E}(0)$ coincides with T_0. This had to be so, because the string does not move at the center. The energy density diverges at the string endpoints $s = \pm\ell/2$. The total energy, however, is finite.

7.5 Motion of closed strings and cusps

Let us now consider the general motion of a free closed string. Just like for open strings, both the wave equation (7.26) and the parameterization constraints (7.27) and (7.28) apply. The wave equation is solved by the expression

$$\vec{X}(t,\sigma) = \frac{1}{2}\big(\vec{F}(u) + \vec{G}(v)\big), \tag{7.65}$$

where we have introduced variables u and v defined by

$$\begin{aligned} u &\equiv ct+\sigma, \\ v &\equiv ct-\sigma. \end{aligned} \tag{7.66}$$

Taking derivatives of $\vec{X}$ we immediately find

$$\frac{1}{c}\frac{\partial\vec{X}}{\partial t} = \frac{1}{2}\big(\vec{F}'(u) + \vec{G}'(v)\big), \tag{7.67}$$

$$\frac{\partial\vec{X}}{\partial\sigma} = \frac{1}{2}\big(\vec{F}'(u) - \vec{G}'(v)\big), \tag{7.68}$$

where primes denote derivatives with respect to argument. We can now form the linear combinations

$$\frac{\partial\vec{X}}{\partial\sigma} + \frac{1}{c}\frac{\partial\vec{X}}{\partial t} = \vec{F}'(u) \quad \text{and} \quad \frac{\partial\vec{X}}{\partial\sigma} - \frac{1}{c}\frac{\partial\vec{X}}{\partial t} = -\vec{G}'(v). \tag{7.69}$$

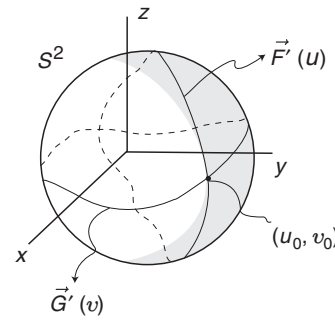

Fig. 7.4 **The tips of the unit vectors $\vec{F}'(u)$ and $\vec{G}'(v)$ define two closed parameterized curves that, in this case, intersect for $u = u_0$ and $v = v_0$.**

The parameterization constraints, summarized in (7.42), then give

$$\left|\vec{F}'(u)\right|^2 = \left|\vec{G}'(v)\right|^2 = 1, \quad \text{for all } u, v. \tag{7.70}$$

For closed strings we do not have boundary conditions but rather a periodicity condition. Since $d\sigma = dE/T_0$ on account of (7.30), the closed string parameter σ has the identification

$$\sigma \sim \sigma + \sigma_1, \quad \text{where} \quad \sigma_1 = E/T_0, \tag{7.71}$$

and E is the energy of the string. As σ increases by σ_1 we are back to the same point on the closed string and therefore

$$\vec{X}(t, \sigma + \sigma_1) = \vec{X}(t, \sigma) \tag{7.72}$$

is the periodicity condition. Making use of (7.65), this condition gives

$$\vec{F}(u + \sigma_1) + \vec{G}(v - \sigma_1) = \vec{F}(u) + \vec{G}(v), \tag{7.73}$$

or, equivalently,

$$\vec{F}(u + \sigma_1) - \vec{F}(u) = \vec{G}(v) - \vec{G}(v - \sigma_1). \tag{7.74}$$

The functions $\vec{F}(u)$ and $\vec{G}(v)$ need not be periodic with period σ_1, but they must change by the same vector when their arguments are increased by σ_1. Since u and v are independent variables, partial derivatives of (7.74) with respect to u and v give

$$\vec{F}'(u + \sigma_1) = \vec{F}'(u) \quad \text{and} \quad \vec{G}'(v + \sigma_1) = \vec{G}'(v). \tag{7.75}$$

Equations (7.70) and (7.75) imply that $\vec{F}'(u)$ and $\vec{G}'(v)$ are periodic unit vectors; they can be described as two independent parameterized closed curves on the surface of a unit two-sphere (Figure 7.4). A closed string motion is fully specified, up to a constant translation, by these two parameterized curves. Indeed, the curves fix $\vec{F}'(u)$ and $\vec{G}'(v)$ and, by integration, $\vec{F}(u)$ and $\vec{G}(v)$ up to integration constants, resulting in $\vec{X}(t, \sigma)$ fixed up to the addition of a constant vector.

An interesting situation arises quite generically. As shown in Figure 7.4, the two parameterized curves $\vec{F}'(u)$ and $\vec{G}'(v)$ may intersect for some values u_0 and v_0 of the parameters u and v:

$$\vec{F}'(u_0) = \vec{G}'(v_0). \tag{7.76}$$

Let t_0 and σ_0 be the values of t and σ defined by u_0 and v_0 through (7.66). It now follows from (7.67) that

$$\frac{1}{c}\frac{\partial \vec{X}}{\partial t}(t_0, \sigma_0) = \frac{1}{2}\big(\vec{F}'(u_0) + \vec{G}'(v_0)\big) = \vec{F}'(u_0). \tag{7.77}$$

Since $\vec{F}'$ is a unit vector we learn that at $t = t_0$ the point $\sigma = \sigma_0$ on the string reaches the speed of light! Moreover, the motion of the point is in the direction of $\vec{F}'(u_0)$. Additional information follows from (7.68):

$$\frac{\partial \vec{X}}{\partial \sigma}(t_0, \sigma_0) = \frac{1}{2}\big(\vec{F}'(u_0) - \vec{G}'(v_0)\big) = \vec{0}. \tag{7.78}$$

This means that the parameterization of the $t = t_0$ string becomes singular at $\sigma = \sigma_0$. To examine the shape of the string near $\sigma = \sigma_0$ we fix $t = t_0$ and use a Taylor expansion around $\sigma = \sigma_0$:

$$\begin{aligned}\vec{X}(t_0, \sigma) = \vec{X}(t_0, \sigma_0) + (\sigma - \sigma_0)\frac{\partial \vec{X}}{\partial \sigma}(t_0, \sigma_0) + \frac{1}{2}(\sigma - \sigma_0)^2\frac{\partial^2 \vec{X}}{\partial \sigma^2}(t_0, \sigma_0) \\ + \frac{1}{3!}(\sigma - \sigma_0)^3\frac{\partial^3 \vec{X}}{\partial \sigma^3}(t_0, \sigma_0) + \cdots.\end{aligned} \tag{7.79}$$

Using (7.78) and the definitions

$$\vec{X}_0 = \vec{X}(t_0, \sigma_0), \quad \vec{T} \equiv \frac{\partial^2 \vec{X}}{\partial \sigma^2}(t_0, \sigma_0), \quad \vec{R} \equiv \frac{\partial^3 \vec{X}}{\partial \sigma^3}(t_0, \sigma_0), \tag{7.80}$$

we find that the expansion (7.79) loses the term linear in $\sigma - \sigma_0$ and becomes

$$\vec{X}(t_0, \sigma) = \vec{X}_0 + \frac{1}{2}(\sigma - \sigma_0)^2\,\vec{T} + \frac{1}{3!}(\sigma - \sigma_0)^3\,\vec{R} + \cdots. \tag{7.81}$$

For generic situations $\vec{T}$ and $\vec{R}$ are nonvanishing and nonparallel. It then follows that for $\sigma = \sigma_0$ we have a *cusp*. A cusp on a string is a point where the two outgoing string directions form zero angle. Equivalently, at a cusp the oriented tangent to the string reverses direction. Equation (7.81) describes a cusp at $\vec{X}_0$: as σ grows from values right below σ_0 to values right above σ_0 the string approaches $\vec{X}_0$ along $\vec{T}$ and recedes from $\vec{X}_0$ along $\vec{T}$. Indeed, near $\vec{X}_0$, $\sigma - \sigma_0$ is very small and the term that contains $\vec{T}$ dominates the expansion (7.81). As we move away from $\vec{X}_0$ the cusp opens up due to the term that contains $\vec{R}$. If we choose a coordinate system in which the tip of the cusp is at the origin, $\vec{T}$ points along the positive y axis, and $\vec{R}$ lies on the (x, y) plane, the cusp traces the curve $y \sim x^{2/3}$, as you will show in Problem 7.7.

If a cusp occurs for (u_0, v_0) it will also occur for $(u_0 + m\sigma_1, v_0 + n\sigma_1)$, with m and n arbitrary integers. As a result, as time goes by, cusps will appear and disappear periodically at various points on the string. Given any two parameterized paths $\vec{F}'(u)$ and $\vec{G}'(v)$ on the

two-sphere, there may be several different intersection points. Each will give rise to a set of cusps.

7.6 Cosmic strings

Our analysis of the classical motion of relativistic strings may be applied to the study of cosmic strings. It is conceivable that physics in the early universe results in strings that are not microscopic but rather expand with the universe to become very large. The motion of cosmic size strings can be study in the classical approximation. As of 2007 no cosmic string has been detected. The discovery of a cosmic string would be a remarkable event. As it turns out, cosmic strings may arise from phenomena unrelated to string theory so, even if discovered, much work will be needed to tell if a cosmic string is a string theory string.

The most direct way to detect a cosmic string is through gravitational lensing. To understand this we begin by discussing the gravitational effects of a straight, infinitely long relativistic string. Suppose you place a massive particle some distance away from such a string. Since the string has rest energy one may imagine that the particle would experience gravitational attraction. This is not the case; the particle would experience exactly zero force. This is a result in general relativity and would not hold in Newtonian gravitation where only the effective mass density μ_0 of the string contributes to the gravitational attraction. In general relativity the tension of the string gives an additional contribution, in fact, a gravitational repulsion. The total attractive force is proportional to $(\mu_0 - \frac{T_0}{c^2})$, a combination that precisely vanishes for relativistic strings.

Although a string does not exert gravitational attraction it affects the geometry of the planes orthogonal to the string. Suppose you go around the string keeping your distance r to the string constant. The circumference $\mathcal{C}$ of the traced circle would be less than the expected value of $2\pi r$. More precisely, for any radial distance r,

$$\frac{\mathcal{C}}{r} = 2\pi - \Delta, \tag{7.82}$$

where the constant Δ is called the *deficit angle*. The two-dimensional spaces orthogonal to the string are in fact cones with deficit angle Δ. The string runs along the apexes of the cones. Indeed cones are spaces in which circular loops at constant distance from the apex satisfy (7.82). This can be made manifest in the construction of a cone starting from the plane. If we represent the plane using the complex variable $z = x + iy$ (as we did when studying orbifolds in Section 2.8) the cone arises by cutting out the region $0 \leq \arg(z) \leq \Delta$ and identifying the resulting boundaries via $z \sim e^{i\Delta} z$. This is shown in Figure 7.5, where we also show a circle with constant distance to the apex $z = 0$. In our discussion of lensing we will assume, for simplicity, that the observer O and the light source S lie on the same cone, namely, they share the same value of the coordinate along the string.

The deficit angle Δ produced by the relativistic string depends on the tension of the string and the value of Newton's constant G. A calculation in general relativity shows that

$$\Delta = \frac{8\pi G T_0}{c^4} = \frac{8\pi G \mu_0}{c^2}. \tag{7.83}$$

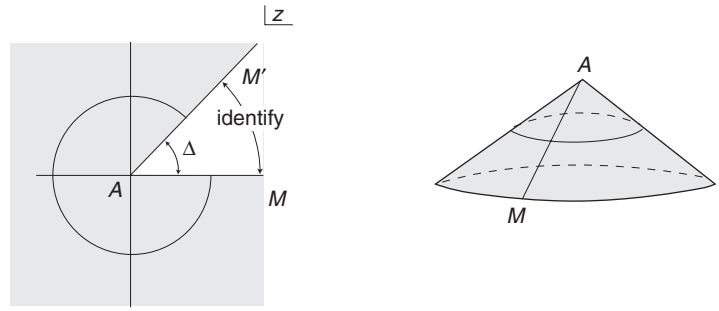

Fig. 7.5 **Left: the complex z plane with the set of points whose argument is greater than zero and less than Δ removed. Right: the familiar picture of a cone obtained by identifying the lines AM and AM' on the left.**

The angle Δ, as we will see soon, is closely related to lensing angles, whose typical values are measured in arc seconds. Noting that $1'' = 4.85 \times 10^{-6}$ rad, we can write

$$\Delta = 5.18'' \times \left(\frac{G\mu_0/c^2}{10^{-6}}\right). \tag{7.84}$$

Quick calculation 7.3 Use the values of G and c to show that

$$\Delta = 3.85'' \times \left(\frac{\mu_0}{10^{21}\text{kg/m}}\right). \tag{7.85}$$

Six kilometers of a string with $\Delta = 3.85''$ would pack a mass equal to that of Earth.

Further insight into the value of Δ can be obtained by writing (7.83) as a manifestly unit-less ratio of the Planck mass m_{P} and a string mass m_s, defined as the unique mass that can be written using only powers of μ_0, c, and $\hbar$.

Quick calculation 7.4 Show that

$$m_s = \sqrt{\frac{\hbar\mu_0}{c}}. \tag{7.86}$$

Recalling that $m_{\text{P}} = \sqrt{\hbar c/G}$ we find

$$\frac{m_s}{m_{\text{P}}} = \sqrt{\frac{G\mu_0}{c^2}}, \tag{7.87}$$

and, as a result,

$$\Delta = 8\pi\left(\frac{m_s}{m_{\text{P}}}\right)^2. \tag{7.88}$$

A small Δ arises for m_s small compared to m_{P}.

To discuss lensing we first review some properties of geodesics. A curve that joins two fixed points is a geodesic if its length is stationary under arbitrary infinitesimal deformations that vanish at the fixed points; the length does not change to first order in such

deformations of the curve. In the plane there is just one geodesic between any two points: the straight line that joins them. This geodesic is also the shortest path between the two points. On more complicated spaces geodesics need not be unique nor have minimum length. There are infinitely many geodesics joining the north pole and the south pole of a two-sphere, each one a half-circle of constant longitude. For two generic points on the two-sphere you get two geodesics, a short one and a long one. Indeed, the two points determine a great circle, and the geodesics are the two complementary arcs bounded by the points. The short geodesic is the shortest path between the two points. The long geodesic is only a path of stationary length, a saddle point for the length function: some deformations make it longer and others make it shorter. In fact, the long geodesic can be continuously deformed into the short geodesic. On a cylinder any two points can be joined by an infinite number of geodesics. The geodesics are characterized by the number of times they wrap around the cylinder while getting from one point to the other.

Quick calculation 7.5 Sketch and find the length of the five shortest geodesics that join the points $(0, 0)$ and $(1, 0)$ on the cylinder $(x, y) \sim (x, y + 1)$.

On a cone there is an intricate pattern of geodesics. Consider two points P and Q and the geodesics from P to Q. The number of geodesics depends on the deficit angle Δ of the cone at the apex A and the angle ϕ between P and Q, defined as the positive angle that by clockwise rotation takes AP to AQ. The angle ϕ must necessarily be smaller than the total angle α at the apex:

$$\phi \leq \alpha \equiv 2\pi - \Delta. \tag{7.89}$$

For a complete calculation of the geodesics see Problem 7.8. Here we restrict ourselves to the case $\Delta \ll 1$, relevant to possible astrophysical situations. Assume that the observer O is a distance d_O from the string and that the light source S is a distance d_S from the string. Lensing occurs if the source S is roughly opposite to O across the string. In this case there are two geodesics that join S and O, one to each side of the string. Light uses both geodesics to reach O, who detects two identical images of the source S. Two identical images is a signature of lensing by strings and occurs because cones have no curvature away from the apex. Lensing by compact objects can produce more than two images and, since the relevant geometry is curved, the images are distorted and can be quite different.

To visualize the geodesics we represent the cone by placing the source on the two radial lines that are to be identified across an angle Δ (Figure 7.6). The source appears as the points S and S', and the two geodesics that join the source and the observer are the straight lines SO and $S'O$. The angle $\delta\phi$ between SO and $S'O$ at O is the lensing angle – the angle between the images seen by the observer. From the figure we see that

$$\delta\phi = \alpha + \beta. \tag{7.90}$$

Moreover, since Δ is the sum of exterior angles for the triangles SAO and $S'AO$, it equals

$$\Delta = \alpha + \beta + \alpha' + \beta'. \tag{7.91}$$

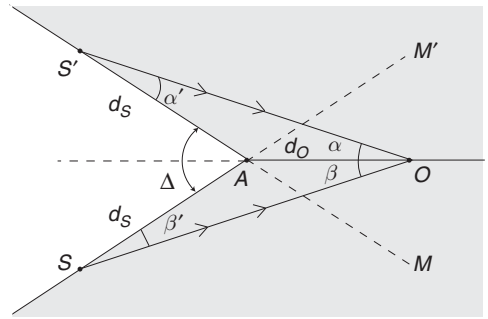

Fig. 7.6 A source S and an obserber O at distances d_S and d_O, respectively, from the apex A of a cone with deficit angle Δ. For simplicity, the cut is along the source, which appears as points S and S' to be identified. Rays SO and $S'O$ from the source reach the observer in directions separated by the lensing angle $\delta\phi = \alpha + \beta$.

Since each angle that appears on the above right-hand side is positive and $\Delta \ll 1$, each one of them is small: $\alpha, \alpha', \beta, \beta' \ll 1$. The law of sines then gives

$$\frac{\sin\alpha}{d_S} = \frac{\sin\alpha'}{d_O}, \quad \frac{\sin\beta}{d_S} = \frac{\sin\beta'}{d_O} \quad \rightarrow \quad \frac{\alpha}{d_S} = \frac{\alpha'}{d_O}, \quad \frac{\beta}{d_S} = \frac{\beta'}{d_O}, \tag{7.92}$$

using the small angle approximation. Solving for α' and β' and substituting back in (7.91), we have

$$\Delta = \alpha + \beta + \frac{d_O}{d_S}(\alpha + \beta) = \delta\phi\Big(1 + \frac{d_O}{d_S}\Big), \tag{7.93}$$

where we also used (7.90). We finally get

$$\delta\phi = \frac{\Delta}{\big(1 + \frac{d_O}{d_S}\big)}. \tag{7.94}$$

We see that the lensing angle $\delta\phi$ is bounded above by Δ. It approaches this upper bound as $d_S \to \infty$. Figure 7.6 also makes clear that lensing occurs only if O lies within the sector MAM'. If O lies outside this sector, there is just one geodesic and O sees a single image.

In the past few years images of galaxies that were candidates for string lensing were later shown, in higher resolution, to be the images of two similar but different galaxies. Since cosmic strings possibly move with relativistic speeds, it may be that lenses exist only for limited periods of time, making their detection challenging.

It may be possible to infer the existence of cosmic strings indirectly. Cosmic strings can give contributions to temperature anisotropies in the cosmic microwave background. The lack of observed evidence to this effect suggests the bound

$$\frac{G\mu_0}{c^2} < 3 \times 10^{-7}. \tag{7.95}$$

Equation (7.84) shows that for strings obeying this bound, lensing angles will be smaller than a couple of arc-seconds. Additionally, the motion of cosmic strings produces gravitational waves. In fact, the cusps we studied in the previous section are efficient generators of gravity waves. It seems possible that such waves could be observed by gravitational wave detectors even for strings of $G\mu_0/c^2 \sim 10^{-13}$. While it remains a long shot, cosmic strings may furnish the first experimental evidence for string theory.

Problems

Problem 7.1 Short proofs concerning open strings with completely free endpoints.

Prove the following four statements that refer to open strings with free endpoints.

(a) If one endpoint of an open string happens to lie at all times on a hyperplane, the full open string lies at all times on the same hyperplane.
(b) If one endpoint of an open string happens to lie at all times within a distance R from a point P_0, the full open string lies at all times within a distance R from P_0.
(c) If one endpoint of an open string happens to lie at all times within a convex subspace, the full open string lies at all times within that convex subspace. [This is the general version of the results proven in (a) and (b).]
(d) Show that the length ℓ of an open string parameterized with energy (Section 7.3) is given by

$$\ell = \int_0^{\sigma_1} \sqrt{1 - \frac{v_\perp^2}{c^2}}\, d\sigma.$$

Problem 7.2 Exploring further the rigidly rotating string.

Let $s \in (-\ell/2, \ell/2)$ be a length parameter on the rigidly rotating string studied in Section 7.4, with $s = 0$ chosen to be the fixed center of the string. Let $\mathcal{E}(s)$ denote the energy per unit length as a function of s.

(a) Show that $\mathcal{E}(s) = T_0/\sqrt{1 - (4s^2/\ell^2)}$. Plot $\mathcal{E}(s)$ as a function of s. Note that $\mathcal{E}(s)$ has integrable singularities at the string endpoints, and confirm that the total energy is $\frac{\pi}{2}\ell T_0$.
(b) For what points on the string is the local energy density equal to the average energy density?
(c) Calculate the energy $\hat{E}(s)$ carried by the string on the interval $[-s, s]$. What is the value of $s/(\ell/2)$ for this energy to be half of the total energy of the string? 90% of the energy?

Problem 7.3 Time evolution of an initially static closed relativistic string.

The time development of closed strings in the static gauge is governed by equations (7.26), (7.27), and (7.28).

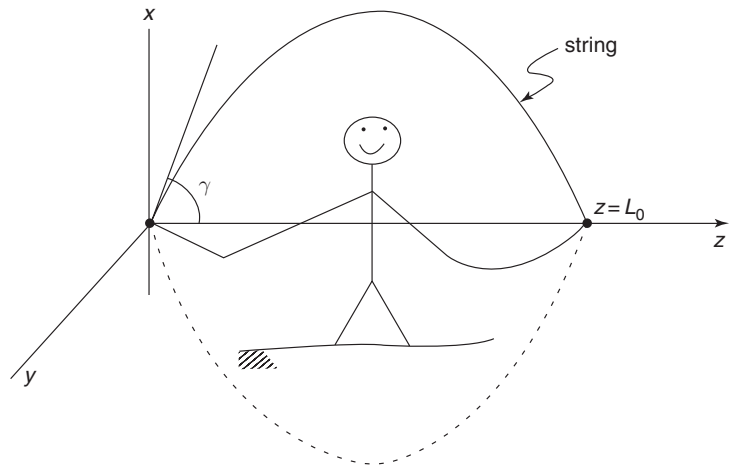

Fig. 7.7 **Problem 7.4: Kasey's relativistic jumping rope.**

(a) Assume that $\frac{\partial \vec{X}}{\partial t}(0, \sigma) = 0$, and write the general solution for $\vec{X}(t, \sigma)$ in terms of a vector function $\vec{F}(u)$ of a single variable. What do the parameterization conditions require on $\vec{F}$?
(b) On a closed string the parameter σ lives on the circle $\sigma \sim \sigma + \sigma_1$. What condition would you impose on $\vec{X}(t, \sigma)$ to implement this feature? What are the implications for $\vec{F}$?
(c) Consider a closed string which at $t = 0$ is static and traces a closed curve γ of length ℓ. Calculate the smallest time $t_P > 0$ that must elapse for the closed string to trace the curve γ again. Is $\vec{X}(t_P, \sigma)$ equal to $\vec{X}(0, \sigma)$?
(d) List the steps you would take with a computer (that can do integrals and invert functions) to produce the time evolution of an initially static closed string of arbitrary shape lying on the (x, y) plane. Assume that the initial string is given to you as the parameterized closed curve $(x(\lambda), y(\lambda))$, with some parameter $\lambda \in [0, \lambda_0]$.

Problem 7.4 Kasey's relativistic jumping rope.

Consider a relativistic open string with fixed endpoints:

$$\vec{X}(t, 0) = \vec{x}_1, \quad \vec{X}(t, \sigma_1) = \vec{x}_2. \tag{1}$$

The boundary condition at $\sigma = 0$ is satisfied by the following solution to the wave equation:

$$\vec{X}(t, \sigma) = \vec{x}_1 + \frac{1}{2}\Big(\vec{F}(ct + \sigma) - \vec{F}(ct - \sigma)\Big). \tag{2}$$

Here $\vec{F}$ is a vector function of a single variable.

(a) Use (2) and the boundary condition at $\sigma = \sigma_1$ to find a condition on $\vec{F}(u)$.
(b) Write down the constraint on $\vec{F}(u)$ that arises from the parameterization conditions (7.42).

As an application, consider Kasey's attempts to use a relativistic open string as a jumping rope. For this purpose, she holds the open string (in three spatial dimensions) with her right hand at the origin $\vec{x}_1 = (0, 0, 0)$ and with her left hand at the point $z = L_0$ on the z axis, or $\vec{x}_2 = (0, 0, L_0)$ (Figure 7.7). As she starts jumping we observe that the tangent vector $\vec{X}'$ to the string at the origin rotates with constant angular velocity around the z axis forming an angle γ with it.

(c) Use the above information to write an expression for $\vec{F}'(u)$.
(d) Find σ_1 in terms of the length L_0 and the angle γ.
(e) Calculate $\vec{X}(t, \sigma)$ for the motion of Kasey's relativistic jumping rope.
(f) How is the energy distributed in the string as a function of z?

Problem 7.5 Planar motion for an open string with fixed endpoints.

Consider the motion of a relativistic open string on the (x, y) plane. The string endpoints are attached to $(x, y) = (0, 0)$ and $(x, y) = (a, 0)$, where $a > 0$. As opposed to the relativistic jumping rope, the string now remains in the (x, y) plane. The motion is described by

$$\vec{X}(t, \sigma) = \frac{1}{2}\Big(\vec{F}(ct + \sigma) - \vec{F}(ct - \sigma)\Big), \tag{1}$$

where $\vec{F}(u)$ is a vector function of a single variable which satisfies

$$\Big|\frac{d\vec{F}}{du}\Big|^2 = 1 \quad \text{and} \quad \vec{F}(u + 2\sigma_1) = \vec{F}(u) + (2a\,,\ 0). \tag{2}$$

Consider an ansatz of the form

$$\vec{F}'(u) \equiv \frac{d\vec{F}}{du} = \Big(\cos\Big[\gamma \cos\frac{\pi u}{\sigma_1}\Big],\ \sin\Big[\gamma \cos\frac{\pi u}{\sigma_1}\Big]\Big). \tag{3}$$

(a) Is this ansatz consistent with the conditions in (2)?
(b) Calculate $\vec{X}'(0, \sigma)$. Letting $\vec{X}(0, \sigma) \equiv (x(\sigma), y(\sigma))$, give $dy/d\sigma$ and plot it as a function of $\sigma \in [0, \sigma_1]$, assuming, for convenience, that $0 < \gamma < \pi/2$. Use this to make a rough sketch of the string position $y(\sigma)$ as a function of σ at $t = 0$.
(c) Calculate $\vec{X}'(t, 0)$ and use it to describe the motion of the string near the origin. What is the interpretation of γ?
(d) Use the second condition in (2) to find an integral relation between a, σ_1 and γ. Assume that γ is small, and find an approximate explicit relation between these three variables keeping terms of order γ^2.
(e) Show that $a/\sigma_1 = J_0(\gamma)$, where J_0 is the Bessel function of order zero. [Hint: look up integral representations of Bessel functions.]

Problem 7.6 Planar motion (*continued*) and the formation of cusps.

We investigate further the solution obtained in Problem 7.5, where it was shown that $a/\sigma_1 = J_0(\gamma)$. Here a is the distance between the fixed endpoints of the open string, and σ_1 is the length parameter of the string: it equals the length of the string at any time when

the string has zero velocity (why?). The relation between a and the angular variable γ is unusual since J_0 is not a periodic function. This means that the open string motions corresponding to γ and $2\pi + \gamma$ are not the same.

(a) Show that the instantaneous slope of the string is described by

$$\vec{X}'(t,\sigma) = \cos\Big(\gamma \, \sin\frac{\pi ct}{\sigma_1} \sin\frac{\pi\sigma}{\sigma_1}\Big) \, (\cos\beta, \sin\beta) \,,$$

where

$$\beta = \gamma \cos\frac{\pi ct}{\sigma_1} \cos\frac{\pi\sigma}{\sigma_1}.$$

Show that at $ct = \sigma_1/2$ the string is horizontal.

(b) Prove that the instantaneous (transverse) velocity of the string satisfies

$$\Big|\frac{1}{c}\frac{\partial\vec{X}}{\partial t}\Big| = \Big|\sin\Big(\gamma \, \sin\frac{\pi ct}{\sigma_1} \sin\frac{\pi\sigma}{\sigma_1}\Big)\Big|.$$

Note that at $t = 0$ the string has zero velocity. Conclude that whenever $\gamma < \pi/2$ no point on the string ever reaches the speed of light. Moreover, show that for $\gamma = \pi/2$ the string midpoint $\sigma = \sigma_1/2$ acquires the speed of light when the string is horizontal.

(c) A tractable case is obtained for $\gamma = \sqrt{2}(\pi/2)$ ($\simeq 127.3°$). Show that at $ct = \sigma_1/4$ one point on the string reaches the speed of light. Examine the string at the slightly later time $ct = \sigma_1/3$, show that there are *two* points that have the speed of light, and find the corresponding values of σ. Analyze $\vec{X}'$ as a function of σ to show that the string has a cusp at each of these points.

(d) Use your favorite mathematical software package to generate the picture of the string considered in (c) at various times (you will use numerical integration). Assume that $a = 1$ and verify that $\sigma_1 \simeq 10.155$. Show the string for $ct = 0,\ \sigma_1/4$, and $\sigma_1/3$.

Problem 7.7 Cusps in the evolution of closed strings.

In this problem we derive a few properties of the cusps that appear generically in the evolution of free closed strings. For this we examine in more detail equation (7.81).

(a) Use the Taylor expansions of $\vec{F}(u)$ and $\vec{G}(v)$ around u_0 and v_0 to prove that

$$\vec{T} = \frac{1}{4}\big(\vec{F}''(u_0) + \vec{G}''(v_0)\big), \quad \vec{R} = \frac{1}{12}\big(\vec{F}'''(u_0) - \vec{G}'''(v_0)\big), \tag{1}$$

where primes denote derivatives with respect to the argument. Assume that the intersection of the paths on the two-sphere indicated in equation (7.76) is regular: the paths are not parallel at the intersection and neither $\vec{F}''(u_0)$ nor $\vec{G}''(v_0)$ vanishes. Explain why $\vec{T}$ is non-zero and orthogonal to $\vec{F}'(u_0)$. In general $\vec{R}$ does not vanish, but it may under special conditions.

(b) Equation (7.81) shows that the cusp opens up along the direction of the vector $\vec{T}$ and, locally, is contained in the plane spanned by $\vec{T}$ and $\vec{R}$. Fix the origin of the coordinate system at $\vec{X}_0$, align the positive y axis along $\vec{T}$, and position the x axis so that $\vec{R}$ lies on the (x, y) plane. Demonstrate that near the cusp $y \sim x^{2/3}$. In what plane does the velocity of the cusp lie?

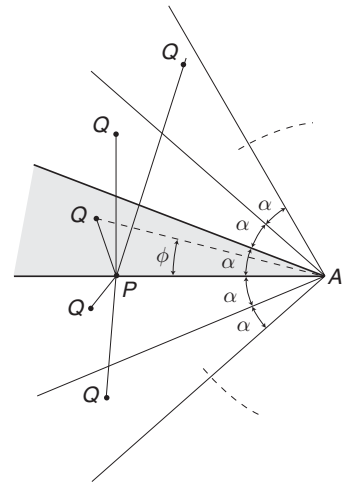

Fig. 7.8 Problem 7.8: a cone, shown shaded, with total internal angle α at the apex and two points P and Q separated by an angle ϕ.

(c) Consider the functions $\vec{F}(u)$ and $\vec{G}(v)$ given by

$$\vec{F}(u) = \frac{\sigma_1}{2\pi}\Big(\sin\frac{2\pi u}{\sigma_1}, -\cos\frac{2\pi u}{\sigma_1}, 0\Big), \quad \vec{G}(v) = \frac{\sigma_1}{4\pi}\Big(\sin\frac{4\pi v}{\sigma_1}, 0, -\cos\frac{4\pi v}{\sigma_1}\Big). \tag{2}$$

Verify that conditions (7.70) and (7.74) are satisfied. For the cusp at $t = \sigma = 0$ give its direction, the plane it lies on, and its velocity. Draw a sketch.

(d) Show that the motion of the closed string has period $\sigma_1/(4c)$. (Remember, as you found in Problem 7.3, that the string returns to its original position in less time than the function $\vec{F}(ct + \sigma)$ takes to repeat itself.) How many cusps are formed during a period?

Problem 7.8 Counting geodesics on a cone.

Let $\Delta < 2\pi$ be the deficit angle of a cone and $\alpha = 2\pi - \Delta$ the angle at the apex A. Consider two points P and Q on the cone separated by an angle $\phi > 0$: the ray from the apex through Q is obtained from the ray from the apex through P by a clockwise rotation through an angle ϕ. The number N of geodesics joining the points P and Q is given by

$$N = \left[\frac{\pi - \phi}{\alpha}\right] + \left[\frac{\pi + \phi}{\alpha}\right] + 1, \tag{1}$$

where $[x]$ denotes the largest integer less than or equal to x. A convenient depiction of the conical region and copies of it is shown in Figure 7.8.

(a) Convince yourself of the validity of (1) by counting the geodesics from P to Q and from P to the images of Q developing in the clockwise direction, as well as those from P to the images of Q developing in the counterclockwise direction.

(b) Verify that N is invariant under $\phi \to \alpha - \phi$ and explain why it should be so.

(c) $\Delta < \pi$ is the case relevant for gravitational lensing. What are the possible values of N as a function of ϕ?

8 World-sheet currents

For physical insight, physicists often turn to ideas of symmetry and invariance. Symmetry properties of dynamical systems and conserved quantities are closely related. We will learn that in string theory there are currents that flow on the two-dimensional world-sheet traced out by the string in spacetime. The conserved charges associated with these currents are key quantities that characterize the free motion of strings. We give a simple physical interpretation of the objects $\mathcal{P}^\tau$ and $\mathcal{P}^\sigma$ which we encountered earlier.

8.1 Electric charge conservation

We begin our study by reviewing the physics and the mathematics of charge conservation in the context of Maxwell theory. This classic example will help us to develop a more general understanding of the concept of conserved currents.

In electromagnetism, the conserved current is the four-vector $j^\alpha = (c\rho, \vec{j})$, where ρ is the electric charge density and $\vec{j}$ is the current density. Why do we say that j^α is a conserved current? By definition, j^α is a conserved current because it satisfies the equation

$$\partial_\alpha j^\alpha = 0\,. \tag{8.1}$$

Any four-vector which satisfies this equation is called a conserved current. The term "conserved current" is a little misleading, but it is a convention. More precisely, we should say that we have a conserved *charge*, because it is really the charge associated with the current that is conserved. Let us see how this conservation arises.

When we separate the space and time indices in (8.1) we get

$$\partial_0 j^0 + \partial_i j^i = \frac{\partial j^0}{\partial x^0} + \nabla \cdot \vec{j} = 0\,. \tag{8.2}$$

Why is this equation a statement of charge conservation? In electromagnetism, the total electric charge $Q(t)$ in a fixed volume V is just the integral of the charge density ρ over the volume:

$$Q(t) = \int_V \rho(t, \vec{x})\, d^3x = \int_V \frac{j^0(t, \vec{x})}{c}\, d^3x. \tag{8.3}$$

Up to a constant, the charge is the integral over space of the first component of the current. Its time derivative is given by

$$\frac{dQ}{dt} = \int_V \frac{\partial j^0}{\partial x^0}\, d^3x. \tag{8.4}$$

Using equation (8.2), we can write

$$\frac{dQ}{dt} = -\int_V \nabla \cdot \vec{j}\, d^3x. \tag{8.5}$$

Letting S denote the boundary of V, the divergence theorem gives

$$\frac{dQ}{dt} = -\int_S \vec{j} \cdot d\vec{a}. \tag{8.6}$$

This equation encodes the physical statement of charge conservation: the charge inside a volume V can only change if there is a flux of current across the surface S which bounds the volume. In many cases, we take V to be so large that the current $\vec{j}$ vanishes on the surface S. In these cases,

$$\frac{dQ}{dt} = 0. \tag{8.7}$$

The charge Q is then time independent and is said to be "conserved." It is also well established that electric charge is a Lorentz invariant: all inertial observers measuring charge obtain the same number. Not all conserved quantities are Lorentz invariant. Energy, for example, is conserved, but energy is *not* Lorentz invariant. These facts are examined in more detail in Problems 8.1 and 8.2.

8.2 Conserved charges from Lagrangian symmetries

One of the most useful properties of Lagrangians is that they can be used to deduce the existence of conserved quantities. Conserved quantities can help us to understand the dynamics of a system. In this section we begin our work in the context of Lagrangian mechanics, learning how to construct the conserved quantity associated with a symmetry. We then turn to Lagrangian densities, and show how to construct the conserved current associated with a symmetry.

Let $L(q(t), \dot{q}(t); t)$ be a Lagrangian that depends on a coordinate $q(t)$, the velocity $\dot{q}(t)$, and may even have explicit time dependence. Consider, moreover, a variation of the coordinate $q(t)$:

$$q(t) \to q(t) + \delta q(t), \tag{8.8}$$

where $\delta q(t)$ is some specific infinitesimal variation. If $q(t)$ represents the path of a particle, for example, the variation above is an instruction on how to change the path: the position at time t is changed by $\delta q(t)$. Suppose that we have a rule that tells us how to vary *any*

path $q(t)$. This means that given any $q(t)$, we know how to construct the corresponding variation $\delta q(t)$. Such a rule can be written in the form

$$\delta q(t) = \epsilon\, h\big(q(t); t\big), \tag{8.9}$$

where ϵ is an infinitesimal constant and h is some function.

As a result of the change (8.8) of the path, there is an induced change in the velocity $\dot{q}(t)$:

$$\dot{q}(t) \to \dot{q}(t) + \frac{d(\delta q(t))}{dt}. \tag{8.10}$$

In order to determine how $L(q(t), \dot{q}(t); t)$ changes as a result of the variation (8.8), we must also vary the velocity $\dot{q}(t)$ according to (8.10). We will generally speak of the variation of $q(t)$ leaving it implicit that the velocity $\dot{q}(t)$ is also varied accordingly. Because δq is infinitesimal, the variation of the Lagrangian consists only of terms which are linear in δq. If those terms vanish, the Lagrangian is said to be invariant. Moreover, the transformation in (8.8) is then said to be a *symmetry transformation*. A rule that tells us how to vary any path in such a way that the Lagrangian is not changed is a symmetry transformation. The rule is specified by the function h in (8.9).

We now state our claim: if the Lagrangian L is invariant under the variation (8.8), then the quantity Q, defined by

$$\epsilon\, Q \equiv \frac{\partial L}{\partial \dot{q}}\, \delta q\,, \tag{8.11}$$

is conserved in time for *physical motion*. That is, for any motion $q(t)$ which satisfies the equations of motion, the "charge" Q is a constant:

$$\frac{dQ}{dt} = 0. \tag{8.12}$$

Note that the ϵ on the left-hand side of (8.11) cancels with the ϵ that appears in δq (see (8.9)).

To prove the conservation of Q, consider the Euler–Lagrange equations that follow from the variation of the action $S = \int dt\, L$ (you may have derived them in Problem 4.8). These equations are:

$$\frac{d}{dt}\Big(\frac{\partial L}{\partial \dot{q}}\Big) - \frac{\partial L}{\partial q} = 0. \tag{8.13}$$

Since the Lagrangian does not change under the coordinate and velocity variations (8.8) and (8.10), we must have

$$\frac{\partial L}{\partial q}\delta q + \frac{\partial L}{\partial \dot{q}}\frac{d}{dt}(\delta q) = 0. \tag{8.14}$$

Using (8.13) to eliminate $\frac{\partial L}{\partial q}$ we obtain

$$\frac{d}{dt}\Big(\frac{\partial L}{\partial \dot{q}}\Big)\delta q + \frac{\partial L}{\partial \dot{q}}\frac{d}{dt}(\delta q) = \frac{d}{dt}\Big(\frac{\partial L}{\partial \dot{q}}\,\delta q\Big) = 0, \tag{8.15}$$

proving that (8.12) holds for Q defined as in (8.11).

Let us apply this new perspective on conservation to a Lagrangian $L(\dot{q}(t))$ that depends only on the velocity. How can we vary $q(t)$ leaving the Lagrangian unchanged? One way is to apply $q(t) \to q(t) + \epsilon$, where ϵ is any constant. In this variation, $\delta q(t) = \epsilon$, and the function h of (8.9) is just equal to one. This is a uniform space translation: at any time t the position of the particle is changed by the same amount ϵ. Correspondingly, the velocity does not change: $\dot{q}(t) \to \dot{q}(t) + d\epsilon/dt = \dot{q}(t)$. Since the Lagrangian only depends on the velocity, it does not change, and we have a symmetry. Making use of (8.11), we find

$$\epsilon Q = \frac{\partial L}{\partial \dot{q}}\delta q = \frac{\partial L}{\partial \dot{q}}\,\epsilon \;\longrightarrow\; Q = \frac{\partial L}{\partial \dot{q}} = p. \tag{8.16}$$

We recognize Q as the momentum associated with the coordinate q. This quantity is conserved because the position q does not appear in the Lagrangian. Note that the conservation equation $dQ/dt = 0$ coincides with the Euler–Lagrange equation (8.13). This example illustrates a familiar result in Lagrangian mechanics: if the Lagrangian of a system does not contain a given coordinate, the momentum conjugate to that coordinate is conserved. For a free nonrelativistic particle, for example, $L = \frac{1}{2}m(\dot{q})^2$. In this case $Q = m\dot{q}$.

Now let us consider Lagrangian *densities* with symmetries. While symmetries of a Lagrangian guarantee the existence of conserved *charges*, symmetries of a Lagrangian density guarantee the existence of conserved *currents*. We write the action as the integral of the Lagrangian density over the full set of coordinates ξ^α of some relevant "world":

$$S = \int d\xi^0 d\xi^1 \ldots d\xi^k \;\mathcal{L}(\phi^a, \partial_\alpha \phi^a). \tag{8.17}$$

Here k denotes the number of space dimensions of the world. The world could be the full Minkowski spacetime, some subspace thereof, or, for example, the two-dimensional parameter space of the string world-sheet. The fields $\phi^a(\xi)$ are functions of the coordinates, and

$$\partial_\alpha \phi^a = \frac{\partial \phi^a}{\partial \xi^\alpha} \tag{8.18}$$

are derivatives of the fields with respect to the coordinates. Each value of the index a corresponds to a field. Consider now the infinitesimal variation

$$\phi^a(\xi) \;\to\; \phi^a(\xi) + \delta\phi^a(\xi), \tag{8.19}$$

and the associated variation for the field derivatives $\partial_\alpha \phi^a$. The infinitesimal variations are conveniently written as the rule

$$\delta\phi^a = \epsilon^i\, h_i^a(\phi), \tag{8.20}$$

that enables us to vary any arbitrary field configuration. Here, ϵ^i are a set of infinitesimal constants, and, for brevity, we omitted all indices in the arguments of h_i^a. We have included an index in ϵ^i because the variation may involve several parameters. A spacetime translation, for example, involves as many parameters as there are spacetime dimensions. Since the index i is repeated in (8.20), it is summed over. You must distinguish clearly the various types of indices we are working with:

$$\begin{aligned} &\alpha && \text{index to label world coordinates } \xi^\alpha, \text{ or vector components,} \\ &i && \text{index to label parameters in the symmetry transformation,} \\ &a && \text{index to label fields in the Lagrangian.} \end{aligned} \tag{8.21}$$

If $\mathcal{L}$ is invariant under (8.19) and the associated variations of the field derivatives, then the quantities j_i^α defined by

$$\epsilon^i \, j_i^\alpha \equiv \frac{\partial \mathcal{L}}{\partial(\partial_\alpha \phi^a)} \delta\phi^a, \tag{8.22}$$

are conserved currents:

$$\partial_\alpha \, j_i^\alpha = 0. \tag{8.23}$$

We will prove this shortly. Equation (8.23) holds for any field configuration that satisfies the Lagrangian equations of motion. In (8.22) the repeated field index a is summed over. If the index i is present in (8.20), then we have *several* currents indexed by i, one for each parameter of the variation. The components of the currents, as many as the number of dimensions in the world, are indexed by α. It is important not to confuse the very different roles that these two kinds of indices play:

$$\begin{aligned} j_i^\alpha\text{:} \quad & i && \text{labels the various currents,} \\ & \alpha && \text{labels the components of the currents.} \end{aligned} \tag{8.24}$$

We showed in Section 8.1 that a conserved current gives rise to a conserved charge. The charge is the integral over space of the zeroth component of the current. It follows that the currents j_i^α give rise to the conserved charges

$$Q_i = \int d\xi^1 d\xi^2 \ldots d\xi^k \; j_i^0. \tag{8.25}$$

We have as many conserved charges as there are parameters in the symmetry transformation.

Quick calculation 8.1 Verify that (8.23) implies that

$$\frac{dQ_i}{d\xi^0} = 0, \tag{8.26}$$

when the currents j_i^α vanish sufficiently rapidly at spatial infinity.

To prove (8.23), we write both the Euler–Lagrange equations associated with the action (8.17) (see Problem 4.8) and the statement of invariance:

$$\partial_\alpha\Big(\frac{\partial\mathcal{L}}{\partial(\partial_\alpha\phi^a)}\Big)-\frac{\partial\mathcal{L}}{\partial\phi^a}=0, \tag{8.27}$$

$$\frac{\partial\mathcal{L}}{\partial\phi^a}\,\delta\phi^a+\frac{\partial\mathcal{L}}{\partial(\partial_\alpha\phi^a)}\partial_\alpha(\delta\phi^a)=0\,. \tag{8.28}$$

Using the first equation to eliminate $\frac{\partial\mathcal{L}}{\partial\phi^a}$ from the second, we get

$$\partial_\alpha\Big(\frac{\partial\mathcal{L}}{\partial(\partial_\alpha\phi^a)}\Big)\,\delta\phi^a+\frac{\partial\mathcal{L}}{\partial(\partial_\alpha\phi^a)}\partial_\alpha(\delta\phi^a)=\partial_\alpha\Big(\frac{\partial\mathcal{L}}{\partial(\partial_\alpha\phi^a)}\,\delta\phi^a\Big)=0\,, \tag{8.29}$$

which demonstrates that (8.23) holds for currents defined as in (8.22).

We will use (8.22) to construct conserved currents which live on the string world-sheet. Actually, conserved charges and conserved currents exist under conditions less stringent than those stated in this section. A transformation can be considered a symmetry if the Lagrangian, or the Lagrangian density, if appropriate, is changed by a total derivative. These ideas are explored in Problem 8.9 and Problem 8.10.

8.3 Conserved currents on the world-sheet

To each string we would like to assign a relativistic momentum p_μ which is conserved if the string is moving freely. Even though the momentum p_μ carries an index, it is not a current, but rather a charge. In fact, since each component of p_μ should be separately conserved, this is a case where we have a set of conserved charges.

In the notation of the previous section, Q_i denote the various charges, so we see that the μ index in p_μ is playing the role of the i index in Q_i; it labels the various charges. What then is the α index in j_i^α? We will see that this index labels the coordinates on the world-sheet. The currents live on the world-sheet!

In the string action (6.39), the Lagrangian density is integrated over the world-sheet coordinates τ and σ, and not over the spacetime coordinates x^μ. In this example, the world of (8.17) is two-dimensional, and the index α in (8.21) takes two values. As a result, the conserved currents live on the world-sheet: they have two components, and they are functions of the world-sheet coordinates. More explicitly,

$$S=\int d\xi^0 d\xi^1\,\mathcal{L}\,(\partial_0 X^\mu,\partial_1 X^\mu),\quad\text{with}\quad(\xi^0,\xi^1)=(\tau,\sigma), \tag{8.30}$$

and $\partial_\alpha=\partial/\partial\xi^\alpha$. Comparing with (8.17), we see that the field variables ϕ^a are simply the string coordinates X^μ. Note that the string action only depends on derivatives of the string coordinates.

To find conserved currents, we need a field variation δX^μ that does not change the Lagrangian density. One such variation is given by

$$\delta X^\mu(\tau, \sigma) = \epsilon^\mu, \tag{8.31}$$

where ϵ^μ is a constant – that is, it does not depend on τ or σ. This transformation is a constant *spacetime* translation: each point on the world-sheet is displaced by the same vector ϵ^μ. The Lagrangian density is invariant because it only depends on derivatives $\partial_\alpha X^\mu$, and their variations vanish: $\delta(\partial_\alpha X^\mu) = \partial_\alpha(\delta X^\mu) = \partial_\alpha \epsilon^\mu = 0$. The role of the various indices in (8.21) is now clear: α is a world-sheet index, i is a spacetime index, and so is a.

Now let us construct the conserved current. Using (8.22), and letting i and a indices run over the values of μ, we have

$$\epsilon^\mu \, j^\alpha_\mu = \frac{\partial \mathcal{L}}{\partial(\partial_\alpha X^\mu)} \, \delta X^\mu = \frac{\partial \mathcal{L}}{\partial(\partial_\alpha X^\mu)} \, \epsilon^\mu \, . \tag{8.32}$$

Cancelling the common ϵ^μ factor on the two sides of this equation, we find an expression for the currents:

$$j^\alpha_\mu = \frac{\partial \mathcal{L}}{\partial(\partial_\alpha X^\mu)} \longrightarrow (\, j^0_\mu, \, j^1_\mu \,) = \Big(\frac{\partial \mathcal{L}}{\partial \dot{X}^\mu}, \, \frac{\partial \mathcal{L}}{\partial X^{\mu\prime}} \Big). \tag{8.33}$$

We have seen such derivatives of $\mathcal{L}$ before; they appeared in equations (6.49) and (6.50). We can in fact identify

$$j^\alpha_\mu = \mathcal{P}^\alpha_\mu \longrightarrow (\, j^0_\mu, \, j^1_\mu \,) = (\, \mathcal{P}^\tau_\mu, \, \mathcal{P}^\sigma_\mu \,). \tag{8.34}$$

This is really interesting: the τ and σ superscripts in $\mathcal{P}_\mu$ label the *components* of a world-sheet current. The equation for current conservation is

$$\partial_\alpha \mathcal{P}^\alpha_\mu = \frac{\partial \mathcal{P}^\tau_\mu}{\partial \tau} + \frac{\partial \mathcal{P}^\sigma_\mu}{\partial \sigma} = 0. \tag{8.35}$$

This is simply the equation of motion (6.53) for the relativistic string.

Since the currents $\mathcal{P}^\alpha_\mu$ are indexed by μ, the conserved charges are also indexed by μ. Following (8.25), to get the charges we integrate the zeroth components $\mathcal{P}^\tau_\mu$ of the currents over space. In the present case, this means integrating over σ:

$$p_\mu(\tau) = \int_0^{\sigma_1} \mathcal{P}^\tau_\mu(\tau, \sigma) \, d\sigma. \tag{8.36}$$

This integral is done with τ held constant. We have called the conserved charges p_μ because they give the spacetime momentum carried by the string. Indeed, we have seen that these charges arise from spacetime translational invariance. Since (8.36) gives the total spacetime momentum as an integral over a string parameterized by σ, we learn that

$\mathcal{P}^\tau_\mu$ is the σ density of spacetime momentum carried by the string. (8.37)

$\mathcal{P}^\tau_\mu$ is the derivative of the Lagrangian density with respect to the velocity $\dot{X}^\mu$, so it has the interpretation of a canonical momentum density, in agreement with (8.37). A similar identification was obtained for nonrelativistic strings: the quantity $\mathcal{P}^t$, which we defined in (4.46), has the interpretation of momentum density on account of (4.43).

To check conservation, we differentiate (8.36) with respect to τ and use the equation of motion (8.35):

$$\frac{dp_\mu}{d\tau} = \int_0^{\sigma_1} \frac{\partial \mathcal{P}^\tau_\mu}{\partial \tau}\, d\sigma = -\int_0^{\sigma_1} \frac{\partial \mathcal{P}^\sigma_\mu}{\partial \sigma}\, d\sigma = -\mathcal{P}^\sigma_\mu \Big|_0^{\sigma_1}. \tag{8.38}$$

For a closed string, the coordinates $\sigma = 0$ and $\sigma = \sigma_1$ represent the same point on the world-sheet, so the right-hand side vanishes. For an open string with free endpoints, the boundary condition (6.56) states that $\mathcal{P}^\sigma_\mu$ vanishes at the endpoints, so, again, the right-hand side vanishes. In both of these cases, p_μ is conserved:

$$\frac{dp_\mu}{d\tau} = 0. \tag{8.39}$$

There is more to this equation than our previous statement (8.12) of charge conservation. The derivative in (8.39) is with respect to τ and *not* with respect to t. So we may ask: is p_μ conserved in world-sheet time or in Minkowski time? We will discuss this question at length in the following section. The short answer is: in both.

For open strings with Dirichlet boundary conditions along some space directions, the momentum carried by the string along those directions may fail to be conserved. Indeed, the boundary condition (6.55) does not guarantee that the right-hand side of (8.38) vanishes. We already noted this possible nonconservation for nonrelativistic strings (Section 4.6). In open string theory, Dirichlet boundary conditions appear when we have D-branes that are not space filling. The momentum of the string can then fail to be conserved, but the total momentum of the string *and* the D-brane is conserved.

8.4 The complete momentum current

Equation (8.39) is very intriguing. It is a conservation law on the world-sheet rather than in spacetime. We could not have expected otherwise – the currents, after all, live on the world-sheet: their indices are world-sheet indices, and their arguments are world-sheet coordinates. As seen in spacetime, the currents vanish everywhere except on the surface traced out by the string.

If we trust the reparameterization invariance of the physics, we can easily obtain a standard *spacetime* conservation law by choosing the static gauge $t = \tau$. The integral in (8.36)

is then an integral over the strings, the lines of constant time as seen by the chosen Lorentz observer. The conservation law (8.39) becomes

$$\frac{dp_\mu}{dt} = 0\,. \tag{8.40}$$

The Lorentz observer confirms that momentum is conserved in *time*.

Equation (8.39), together with (8.36), tells us that for any fixed world-sheet parameterization, we can compute a unique quantity p_μ using any line of constant τ. This quantity must coincide with the time-independent momentum p_μ obtained using the static gauge, as we now explain.

Consider a moving string, a fixed Lorentz observer, and a particular choice of parameterization on the world-sheet. In this parameterization, over some region of the world-sheet the lines of constant τ are lines of constant t. As a result, over this region we have the static gauge. Over the rest of the world-sheet the parameterization changes smoothly: the lines of constant τ are no longer lines of constant time. On the static gauge region the integral (8.36) gives us the time-independent momentum carried by the strings. Because of (8.39), on the rest of the world-sheet the integral (8.36) must still give the same value for p_μ, even though the lines of constant τ are not strings.

This argument suggests that *any* curve on the world-sheet can be used to calculate the conserved momentum p_μ. Equation (8.36), however, only tells us how to calculate the momentum if the curve is a curve of constant τ. We now show how to generalize (8.36) to be able to compute the momentum p_μ using an (almost) arbitrary curve on the world-sheet *together* with an arbitrary parameterization of the world-sheet. When we deal with open strings, the curve must stretch from one boundary to the other. When we deal with closed strings, the curve must be a closed noncontractible curve.

Reconsider equation (8.36), which integrates the τ-component of the two-dimensional current $(\mathcal{P}^\tau_\mu,\ \mathcal{P}^\sigma_\mu)$ over a curve of constant τ. The quantity that this integral computes is actually the *flux* of the current across the curve. Since the σ-component $\mathcal{P}^\sigma_\mu$ of the current is tangent to the curve of constant τ, it does not contribute to the flux. More generally, consider an infinitesimal segment $(d\tau, d\sigma)$ along an oriented closed curve Γ that encloses a simply connected region $\mathcal{R}$ of the world-sheet (Figure 8.1). Since $(d\tau, d\sigma)$ is parallel to the oriented tangent, the outgoing normal to the segment is $(d\sigma, -d\tau)$. It is reasonable to define the outgoing flux of the current across the segment as the scalar product of the current vector and the outgoing normal vector:

$$\text{Infinitesimal flux} = (\mathcal{P}^\tau_\mu,\ \mathcal{P}^\sigma_\mu)\cdot(d\sigma, -d\tau) = \mathcal{P}^\tau_\mu\, d\sigma - \mathcal{P}^\sigma_\mu\, d\tau. \tag{8.41}$$

We now show that the flux, so defined, vanishes when computed across a contractible closed curve Γ on the world-sheet. This is a reasonable result because a contractible curve

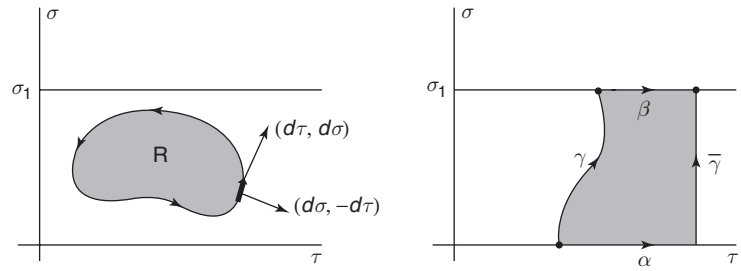

Fig. 8.1 **Left: the total momentum flux out of a simply connected region $\mathcal{R}$ of the world-sheet is zero. Right: a simply connected region whose boundary includes an arbitrary curve γ and a string $\bar{\gamma}$.**

surrounds a domain $\mathcal{R}$, and we do not expect a domain to be a momentum source or sink. The outgoing flux across Γ is written as

$$p_\mu(\Gamma) = \oint_\Gamma \left(\mathcal{P}^\tau_\mu \, d\sigma - \mathcal{P}^\sigma_\mu \, d\tau\right). \tag{8.42}$$

By the two-dimensional version of the divergence theorem, the flux of the current $\mathcal{P}^\alpha_\mu$ out of $\mathcal{R}$ is equal to the integral of the divergence of $\mathcal{P}^\alpha_\mu$ over $\mathcal{R}$:

$$p_\mu(\Gamma) = \int_{\mathcal{R}} \left(\frac{\partial \mathcal{P}^\tau_\mu}{\partial \tau} + \frac{\partial \mathcal{P}^\sigma_\mu}{\partial \sigma}\right) d\tau d\sigma = 0, \tag{8.43}$$

since $\mathcal{P}^\alpha_\mu$ is a conserved current. This is what we wanted to show.

Quick calculation 8.2 Consider $\mathbb{R}^2$ with coordinates (x, y) and a simply connected region M surrounded by a boundary Γ with counterclockwise orientation. The divergence theorem for a vector (A^x, A^y) reads

$$\oint_\Gamma (A^x dy - A^y dx) = \iint_M \left(\frac{\partial A^x}{\partial x} + \frac{\partial A^y}{\partial y}\right) dx dy. \tag{8.44}$$

Verify that this equation holds for a tiny rectangle with corners (x_0, y_0), $(x_0 + dx, y_0)$, $(x_0 + dx, y_0 + dy)$, and $(x_0, y_0 + dy)$. This suffices to prove (8.44) by breaking M into a collection of tiny rectangles. We used (8.44), for an $\mathbb{R}^2$ with coordinates (τ, σ) and a vector $(\mathcal{P}^\tau_\mu, \mathcal{P}^\sigma_\mu)$, to pass from (8.42) to (8.43).

We now generalize (8.36) as follows. For an arbitrary curve γ that starts on the $\sigma = 0$ boundary of the world-sheet and ends on the $\sigma = \sigma_1$ boundary, we define

$$p_\mu(\gamma) = \int_\gamma \left(\mathcal{P}^\tau_\mu \, d\sigma - \mathcal{P}^\sigma_\mu \, d\tau\right). \tag{8.45}$$

If γ is a curve of constant τ, then $d\tau = 0$ all along γ, and $p_\mu(\gamma)$ reduces to (8.36). We now prove that $p_\mu(\gamma)$, as defined in all generality by (8.45), actually coincides with p_μ, as defined in (8.36). Consider a curve γ that stretches from one side of the world-sheet to the other, and a curve $\bar{\gamma}$ of constant τ, as shown to the right of Figure 8.1. Let α and β denote

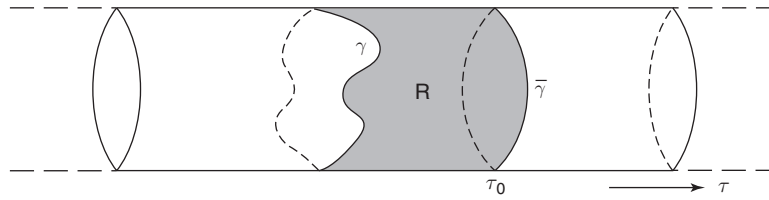

Fig. 8.2 A closed string world-sheet with an arbitrary nontrivial closed curve γ and a closed curve $\bar{\gamma}$ at constant τ. The region between the curves is $\mathcal{R}$.

oriented paths along the world-sheet boundary such that the closed curve Γ surrounding the shaded region is given by

$$\Gamma = \bar{\gamma} - \beta - \gamma + \alpha. \tag{8.46}$$

The curve Γ is oriented counterclockwise and it is contractible. As a result, the flux $p_\mu(\Gamma)$ vanishes:

$$p_\mu(\Gamma) = \int_\Gamma \left(\mathcal{P}^\tau_\mu\, d\sigma - \mathcal{P}^\sigma_\mu\, d\tau\right) = \left(\int_{\bar{\gamma}} - \int_\gamma + \int_\alpha - \int_\beta\right)\left(\mathcal{P}^\tau_\mu\, d\sigma - \mathcal{P}^\sigma_\mu\, d\tau\right) = 0. \tag{8.47}$$

Since α and β are curves where $d\sigma$ vanishes, only $\mathcal{P}^\sigma_\mu\, d\tau$ contributes to the integrals. But $\mathcal{P}^\sigma_\mu$ vanishes at the string endpoints (for free endpoints), so these integrals vanish identically. Only the integrals over γ and $\bar{\gamma}$ survive, so we have

$$\int_\gamma \left(\mathcal{P}^\tau_\mu\, d\sigma - \mathcal{P}^\sigma_\mu\, d\tau\right) = \int_{\bar{\gamma}} \left(\mathcal{P}^\tau_\mu\, d\sigma - \mathcal{P}^\sigma_\mu\, d\tau\right) = \int_{\bar{\gamma}} \mathcal{P}^\tau_\mu\, d\sigma = p_\mu, \tag{8.48}$$

where we noted that $d\tau = 0$ on $\bar{\gamma}$ and used (8.36). This proves that $p_\mu(\gamma) = p_\mu$ for any contour γ which connects the $\sigma = 0$ and $\sigma = \sigma_1$ boundaries of the world-sheet. We can thus rewrite (8.45) as

$$p_\mu = \int_\gamma \left(\mathcal{P}^\tau_\mu\, d\sigma - \mathcal{P}^\sigma_\mu\, d\tau\right). \tag{8.49}$$

Conservation is now the statement that the integral above is independent of the chosen curve γ, as long as the endpoints of γ lie on the boundary components of the world-sheet.

The arguments are similar in the case of closed strings. We consider an arbitrary nontrivial closed curve γ winding once around the world-sheet, and another similarly nontrivial closed curve $\bar{\gamma}$ of constant τ, for some arbitrary, but fixed, parameterization. The two curves form the boundary of an annular region $\mathcal{R}$ (Figure 8.2). A completely analogous argument shows that both contours give the same result for p_μ. Therefore, we can calculate the momentum of the closed string using any closed curve that winds once around the world-sheet.

How does an arbitrary Lorentz observer use equation (8.49)? The observer looks at a string at some time t, and asks for its momentum. This requires the use of (8.49), with a curve γ

which corresponds to the string in question. With an arbitrary parameterization γ need not be a constant τ curve. At some later time t', the observer asks again for the momentum. At this time, the string corresponds to a curve γ', generically different from γ. By the curve independence of (8.49), the observer concludes that the momentum did not change. The momentum was conserved in time.

We conclude this section by making contact with familiar results. Using a constant τ string γ to evaluate (8.49) we have

$$p^0 = \int_\gamma \mathcal{P}^{\tau 0}\, d\sigma, \quad \vec{p} = \int_\gamma \vec{\mathcal{P}}^{\tau}\, d\sigma. \tag{8.50}$$

The values of the momentum densities in static gauge can be read from (6.94). We thus find that the energy and the spatial momentum of the string are given by:

$$p^0 \equiv \frac{E}{c} = \frac{1}{c}\int_\gamma \frac{T_0 ds}{\sqrt{1-\frac{v_\perp^2}{c^2}}}, \quad \vec{p} = \int_\gamma \frac{T_0 ds}{c^2}\frac{\vec{v}_\perp}{\sqrt{1-\frac{v_\perp^2}{c^2}}}. \tag{8.51}$$

The above equation for the energy of the string coincides with (7.9). The expression for the momentum is reasonable: the momentum carried by a little piece of string is given by its rest mass $T_0 ds/c^2$ times its velocity, times its relativistic factor γ.

8.5 Lorentz symmetry and associated currents

By construction, the action of the relativistic string is Lorentz invariant. It is written in terms of Lorentz vectors that are contracted to build Lorentz scalars. Concretely, this means that Lorentz transformations of the coordinates X^μ leave the action invariant. In this section we will construct the conserved charges associated with Lorentz symmetry.

These charges will be particularly useful when we study quantum string theory in Chapter 12. Whenever we quantize a classical system, there is the possibility that crucial symmetries of the classical theory will be lost. If Lorentz invariance were lost upon quantization, quantum string theory would be very problematic, to say the least. We will have to make sure that the quantum theory is Lorentz invariant. To calculate the conserved charges, we first need the infinitesimal form of the general Lorentz transformations. We recall (Section 2.2) that Lorentz transformations are linear transformations of the coordinates X^μ that leave the quadratic form $\eta_{\mu\nu}X^\mu X^\nu$ invariant. Every infinitesimal linear transformation is of the form $X^\mu \to X^\mu + \delta X^\mu$, where

$$\delta X^\mu = \epsilon^{\mu\nu} X_\nu\,. \tag{8.52}$$

Here $\epsilon^{\mu\nu}$ is a matrix of infinitesimal constants. Lorentz invariance imposes conditions on the constants $\epsilon^{\mu\nu}$. We require $\delta(\eta_{\mu\nu}X^\mu X^\nu) = 0$, and therefore

$$2\eta_{\mu\nu}(\delta X^\mu)X^\nu = 2\eta_{\mu\nu}(\epsilon^{\mu\rho}X_\rho)\,X^\nu = 2\,\epsilon^{\mu\rho}X_\rho X_\mu = 0. \tag{8.53}$$

Imagine decomposing the matrix ϵ into its antisymmetric part and its symmetric part. The antisymmetric part would not contribute to $\epsilon^{\mu\rho}X_\rho X_\mu$. The vanishing of $\epsilon^{\mu\rho}X_\rho X_\mu$, for

all values of X_μ, implies that the symmetric part of ϵ is zero. It follows that the general solution is represented by an antisymmetric $\epsilon^{\mu\nu}$:

$$\epsilon^{\mu\nu} = -\epsilon^{\nu\mu}. \tag{8.54}$$

Infinitesimal Lorentz transformations are thus very simple: they are transformations of the form $\delta X^\mu = \epsilon^{\mu\nu} X_\nu$, with $\epsilon^{\mu\nu}$ antisymmetric. This result applies for any number of spacetime dimensions since Lorentz transformations always leave $\eta_{\mu\nu} X^\mu X^\nu$ invariant.

Quick calculation 8.3 Consider a fixed 2-by-2 matrix A^{ab} $(a, b = 1, 2)$ that satisfies $A^{ab} v_a v_b = 0$ for *all* values of v_1 and v_2. Write out the four terms on the left-hand side of the vanishing condition, and confirm explicitly that the matrix A^{ab} must be antisymmetric.

Quick calculation 8.4 Repeat the above exercise for a 4-by-4 matrix $\epsilon^{\mu\nu}$ $(\mu, \nu = 0, 1, 2, 3)$ that satisfies $\epsilon^{\mu\nu} v_\mu v_\nu = 0$ for *all* values of v_0, v_1, v_2, and v_3.

Quick calculation 8.5 Show that (8.54) implies $\epsilon_{\mu\nu} = -\epsilon_{\nu\mu}$.

Quick calculation 8.6 Examine the boost in (2.36) for very small β. Write $x'^\mu = x^\mu + \epsilon^{\mu\nu} x_\nu$ and calculate the entries of the matrix $\epsilon^{\mu\nu}$. Show that $\epsilon^{10} = -\epsilon^{01} = \beta$, and that all other entries are zero. This confirms that $\epsilon^{\mu\nu}$ is antisymmetric for an infinitesimal boost.

Let us now show explicitly that the string Lagrangian density is invariant under Lorentz transformations. All terms that appear in this density are of the form

$$\eta_{\mu\nu} \frac{\partial X^\mu}{\partial \xi^\alpha} \frac{\partial X^\nu}{\partial \xi^\beta}, \tag{8.55}$$

where ξ^α and ξ^β are either τ or σ. We claim that any such term is Lorentz invariant. Indeed,

$$\begin{aligned} \delta\Big(\eta_{\mu\nu} \frac{\partial X^\mu}{\partial \xi^\alpha} \frac{\partial X^\nu}{\partial \xi^\beta}\Big) &= \eta_{\mu\nu}\Big(\frac{\partial \delta X^\mu}{\partial \xi^\alpha} \frac{\partial X^\nu}{\partial \xi^\beta} + \frac{\partial X^\mu}{\partial \xi^\alpha} \frac{\partial \delta X^\nu}{\partial \xi^\beta}\Big) \\ &= \eta_{\mu\nu}\Big(\epsilon^{\mu\rho} \frac{\partial X_\rho}{\partial \xi^\alpha} \frac{\partial X^\nu}{\partial \xi^\beta} + \epsilon^{\nu\rho} \frac{\partial X^\mu}{\partial \xi^\alpha} \frac{\partial X_\rho}{\partial \xi^\beta}\Big) \\ &= \epsilon_{\nu\rho} \frac{\partial X^\rho}{\partial \xi^\alpha} \frac{\partial X^\nu}{\partial \xi^\beta} + \epsilon_{\mu\rho} \frac{\partial X^\mu}{\partial \xi^\alpha} \frac{\partial X^\rho}{\partial \xi^\beta}, \end{aligned} \tag{8.56}$$

where we used η to lower the first index on the ϵ constants. Letting $\mu \to \rho$ and $\rho \to \nu$ in the second term we get

$$\delta\Big(\eta_{\mu\nu} \frac{\partial X^\mu}{\partial \xi^\alpha} \frac{\partial X^\nu}{\partial \xi^\beta}\Big) = (\epsilon_{\nu\rho} + \epsilon_{\rho\nu}) \frac{\partial X^\rho}{\partial \xi^\alpha} \frac{\partial X^\nu}{\partial \xi^\beta} = 0, \tag{8.57}$$

because of the antisymmetry of ϵ. This explicitly proves the Lorentz invariance of the string action.

We can now use equation (8.22) to write the currents. It follows from (8.52) that the role of the small parameter ϵ^i is played by $\epsilon^{\mu\nu}$. We therefore have

$$\epsilon^{\mu\nu} j^\alpha_{\mu\nu} = \frac{\partial \mathcal{L}}{\partial(\partial_\alpha X^\mu)} \delta X^\mu = \mathcal{P}^\alpha_\mu \epsilon^{\mu\nu} X_\nu. \tag{8.58}$$

The current $j^{\alpha}_{\mu\nu}$, since it is multiplied by the antisymmetric matrix $\epsilon^{\mu\nu}$, can be defined to be antisymmetric – any symmetric part would drop out of the left-hand side. Using the antisymmetry of $\epsilon^{\mu\nu}$, the right-hand side is written as

$$\epsilon^{\mu\nu}\, j^{\alpha}_{\mu\nu} = (-\tfrac{1}{2}\epsilon^{\mu\nu})\left(X_\mu \mathcal{P}^{\alpha}_{\nu} - X_\nu \mathcal{P}^{\alpha}_{\mu}\right). \tag{8.59}$$

The currents can be read directly from this equation because the factor multiplying $\epsilon^{\mu\nu}$ on the right-hand side is explicitly antisymmetric. Since the overall normalization of the currents is for us to choose, we define the currents $\mathcal{M}^{\alpha}_{\mu\nu}$ by

$$\mathcal{M}^{\alpha}_{\mu\nu} = X_\mu\, \mathcal{P}^{\alpha}_{\nu} - X_\nu\, \mathcal{P}^{\alpha}_{\mu}. \tag{8.60}$$

By construction,

$$\mathcal{M}^{\alpha}_{\mu\nu} = -\mathcal{M}^{\alpha}_{\nu\mu}. \tag{8.61}$$

The equation of current conservation is

$$\frac{\partial \mathcal{M}^{\tau}_{\mu\nu}}{\partial \tau} + \frac{\partial \mathcal{M}^{\sigma}_{\mu\nu}}{\partial \sigma} = 0, \tag{8.62}$$

and the charges, in analogy with (8.49), are given by

$$M_{\mu\nu} = \int_{\gamma} \left(\mathcal{M}^{\tau}_{\mu\nu} d\sigma - \mathcal{M}^{\sigma}_{\mu\nu} d\tau\right). \tag{8.63}$$

The charges, just like the currents, are antisymmetric:

$$M_{\mu\nu} = -M_{\nu\mu}. \tag{8.64}$$

The conservation of $M_{\mu\nu}$ is a result of the contour independence of the definition (8.63). For closed strings, contour independence is guaranteed by the argument given earlier in the case of momentum charges. For free open strings, one point must be addressed to ensure contour independence. The $M_{\mu\nu}$ integrals must receive no contributions from lines on the boundary of the world-sheet. This, in turn, requires the vanishing of $\mathcal{M}^{\sigma}_{\mu\nu}$ on the boundary. This condition is satisfied because $\mathcal{M}^{\sigma}_{\mu\nu}$ involves $\mathcal{P}^{\sigma}$ multiplicatively, and $\mathcal{P}^{\sigma}$ vanishes on the world-sheet boundary. As explained for the case of the momentum charges, a Lorentz observer measuring $M_{\mu\nu}$ using strings at different times will conclude that $dM_{\mu\nu}/dt = 0$.

We can also compute the Lorentz charges $M_{\mu\nu}$ using constant τ lines. In that case,

$$M_{\mu\nu} = \int \mathcal{M}^{\tau}_{\mu\nu}(\tau,\sigma)\, d\sigma = \int (X_\mu \mathcal{P}^{\tau}_{\nu} - X_\nu \mathcal{P}^{\tau}_{\mu})\, d\sigma. \tag{8.65}$$

Since the $M_{\mu\nu}$ are antisymmetric, we have six conserved charges in four dimensions. Letting i and j denote space indices, M_{0i} are the three charges associated with the three basic boosts, and M_{ij} are the three charges associated with the three basic rotations. Given that $\vec{\mathcal{P}}^{\tau}$ is the momentum density, the normalization chosen in (8.60) ensures that $\mathcal{M}^{\tau}_{ij}$ is precisely the angular momentum density. As a consequence, the components M_{ij} measure the string angular momentum $\vec{L}$ via the usual relations $L_i = \frac{1}{2}\epsilon_{ijk}M_{jk}$. Here ϵ_{ijk} is the totally antisymmetric symbol which satisfies $\epsilon_{123} = 1$. More explicitly, we have $L_1 = M_{23}$, $L_2 = M_{31}$, and $L_3 = M_{12}$.

The charges associated with the boosts are:

$$M^{0i} = \int d\sigma \left(ct\, \mathcal{P}^{\tau i} - X^i \mathcal{P}^{\tau 0}\right) = ct\; p^i - \int d\sigma\, X^i \mathcal{P}^{\tau 0}. \tag{8.66}$$

Multiplying by c/E, where E is the conserved energy of the string,

$$\frac{cM^{0i}}{E} = t\,\frac{c^2 p^i}{E} - \frac{1}{E}\int d\sigma\, X^i\, c\mathcal{P}^{\tau 0}. \tag{8.67}$$

Since $c\mathcal{P}^{\tau 0}$ is the energy density along the string, the last term on the above right-hand side can be identified with the time-dependent position $X^i_{\rm cm}(t)$ of the center of mass (energy) of the string. We thus obtain,

$$X^i_{\rm cm}(t) = -\frac{cM^{0i}}{E} + t\,\frac{c^2 p^i}{E}. \tag{8.68}$$

The quantity $c^2 p^i/E$ has the interpretation of center of mass velocity because it coincides with the velocity of a point particle with momentum p^i and energy E. Equation (8.68) describes the motion of the center of mass. The conserved charges M^{0i}, together with E, determine the $t = 0$ position of the center of mass.

8.6 The slope parameter α'

The string tension T_0 is the only dimensionful parameter in the string action. In this section we will motivate the definition of an alternative parameter: the slope parameter α'. These two parameters are related, we can use one or the other. The parameter α' has an interesting physical interpretation, used since the early days of string theory. If we consider a rigidly rotating open string, α' is the proportionality constant that relates the angular momentum J of the string, measured in units of $\hbar$, to the square of its energy E. More explicitly,

$$\frac{J}{\hbar} = \alpha' E^2. \tag{8.69}$$

Since the left-hand side has no units, the units of α' are those of inverse energy-squared:

$$[\alpha'] = \frac{1}{[E]^2}. \tag{8.70}$$

The appearance of $\hbar$ in (8.69) is just a convention. The constant α' was introduced in the quantum theory of strings, and its relation to the string tension involves $\hbar$. For our present purposes, however, the important fact is just the proportionality between J and E^2. This relation does *not* involve $\hbar$ when we use the string tension T_0.

To verify the proportionality implied in equation (8.69), we consider a straight open string of energy E which rotates rigidly on the (x, y) plane. This is precisely the problem examined in Section 7.4. The only nonvanishing component of angular momentum is M_{12}, and its magnitude is denoted by $J = |M_{12}|$. Equation (8.65) tells us that

$$M_{12} = \int_0^{\sigma_1} (X_1 \mathcal{P}_2^\tau - X_2 \mathcal{P}_1^\tau)\, d\sigma. \tag{8.71}$$

To evaluate this integral we need formulae for the position and momenta of the rotating string. We recall equation (7.63),

$$\vec{X}(t, \sigma) = \frac{\sigma_1}{\pi} \cos\frac{\pi\sigma}{\sigma_1} \left(\cos\frac{\pi c t}{\sigma_1},\ \sin\frac{\pi c t}{\sigma_1} \right), \tag{8.72}$$

which records the components (X_1, X_2) of the rotating string. Using equation (7.31), we find

$$\vec{\mathcal{P}}^\tau = \frac{T_0}{c^2} \frac{\partial \vec{X}}{\partial t} = \frac{T_0}{c} \cos\frac{\pi\sigma}{\sigma_1} \left(-\sin\frac{\pi c t}{\sigma_1},\ \cos\frac{\pi c t}{\sigma_1} \right), \tag{8.73}$$

where the right-hand side gives the components $(\mathcal{P}_1^\tau, \mathcal{P}_2^\tau)$. The integral in (8.71) is thus given by

$$M_{12} = \frac{\sigma_1}{\pi} \frac{T_0}{c} \int_0^{\sigma_1} \cos^2\frac{\pi\sigma}{\sigma_1}\, d\sigma = \frac{\sigma_1^2 T_0}{2\pi c}. \tag{8.74}$$

The time dependence disappeared, as expected for a conserved charge. Since $J = |M_{12}|$ and $\sigma_1 = E/T_0$, we have found that

$$J = \frac{1}{2\pi T_0 c} E^2. \tag{8.75}$$

As anticipated, the angular momentum is proportional to the square of the energy of the string. Comparing with equation (8.69) we deduce that

$$\alpha' = \frac{1}{2\pi T_0 \hbar c} \quad \text{and} \quad T_0 = \frac{1}{2\pi\alpha' \hbar c}. \tag{8.76}$$

These equations relate the slope parameter α' to the string tension T_0.

Quick calculation 8.7 To appreciate how unusual the relation $J \sim E^2$ is, consider a one-dimensional straight bar of fixed length and fixed (uniform) mass that is rotating around its midpoint. Show that the nonrelativistic energy and the angular momentum are related by $J \sim \sqrt{E}$.

One may have anticipated the relationship $J \sim E^2$ by the following admittedly rough argument. We know that for rigid rotations $J = I\omega$, where I is the moment of inertia. For an object of mass M and length scale L, we have $I \sim ML^2$. So, $J \sim ML^2\omega$. For a rotating relativistic string $M \sim E$, $L \sim E$, and $\omega \sim 1/E$ (see (7.57)). As a result $J \sim E^2$.

The name slope parameter arises because α' is the slope of the lines of $J/\hbar$, when plotted as a function of energy-squared. In fact, Regge trajectories are approximate lines that arise when plotting angular momentum as a function of energy-squared for hadronic excitations. In the early 1970s, when string theory was investigated as a theory of strong interactions, the slope parameter α' was the experimentally measured quantity that entered into the string action. The action (6.39), which we wrote in terms of T_0, takes the form

$$S = -\frac{1}{2\pi\alpha'\,\hbar c^2}\int_{\tau_i}^{\tau_f} d\tau \int_0^{\sigma_1} d\sigma \sqrt{(\dot{X}\cdot X')^2 - (\dot{X})^2(X')^2}\,. \tag{8.77}$$

It turns out that most of the modern work on string theory uses the slope parameter α' as opposed to the string tension T_0. In Section 3.6 we used $\hbar$, c, and Newton's constant G to calculate a characteristic length $\ell_{\rm P}$ called the Planck length. In string theory, we can use $\hbar$, c, and the dimensionful parameter α' to construct a characteristic length ℓ_s called the string length.

Quick calculation 8.8 Show that

$$\ell_s = \hbar c\,\sqrt{\alpha'}. \tag{8.78}$$

Up to factors of $\hbar$ and c, the string length ℓ_s is the square root of α'. This connection to a fundamental length scale provides an alternative physical interpretation for the slope parameter α'.

Problems

Problem 8.1 Lorentz invariants and their densities.

We want to understand how Lorentz invariance of electric charge implies that charge *density* is the zeroth component of a Lorentz vector.

Consider two Lorentz frames S and S' with S' boosted along the $+x$ direction of S with velocity v. In frame S we have a cubic box of volume L^3 with edges, all of length L, aligned with the axes of the coordinate system. The box is at rest and is filled by a material, also at rest, that has uniform charge density ρ_0. In this frame the current density vanishes $\vec{j} = \vec{0}$. Note that S' sees the box Lorentz contracted along the x direction.

(a) Use the Lorentz invariance of charge to calculate the charge density and current density $(c\rho', \vec{j}')$ that describe the material in the box according to frame S'. Verify that the values $(c\rho_0, \vec{0})$ and $(c\rho', \vec{j}')$ behave as four-vectors $j^\mu \equiv (c\rho_0, \vec{0})$ and $j'^\mu \equiv (c\rho', \vec{j}')$,

in the sense that the quantities $(c\rho', \vec{j}')$ are obtained from the rest frame values $(c\rho_0, \vec{0})$ by a Lorentz boost.

(b) Write a formula for the current j^μ in terms of ρ_0 and the four-velocity of the box, and check that your result works both for frames S and S'.

Problem 8.2 Lorentz invariance of electric charge.

The explicit test of Lorentz invariance of electric charge is quite subtle, so this is a *challenging* problem! We consider, as usual, a frame S and a frame S' moving along the $+x$ direction of S with velocity v. In frame S the total charge is given by integration over all of space at a fixed time $t = 0$:

$$Q \equiv \int d^3x\, \rho(t{=}0, \vec{x}),$$

and in the primed frame S' it is given by an integral over all of space at $t' = 0$:

$$Q' \equiv \int d^3x'\, \rho'(t'{=}0, \vec{x}').$$

We want to prove that $Q' = Q$. The complication is that S and S' do not agree on what is equal time. We will assume that $j^\alpha = (c\rho, \vec{j})$ is a four-vector and it is conserved: $\partial_\alpha j^\alpha = 0$. Moreover, charge densities and currents are taken to vanish at infinity for all times.

(a) Write explicitly the transformations that relate $(t, \vec{x})$ to $(t', \vec{x}')$ and $(c\rho, \vec{j})$ to $(c\rho', \vec{j}')$. Think of ρ and $\vec{j}$ as functions of arguments (t, x, y, z). Prove that

$$Q' = \int d^3x'\, \gamma \left[\rho\left(\gamma \frac{vx'}{c^2}, \gamma x', y', z'\right) - \frac{v}{c^2} j^x \left(\gamma \frac{vx'}{c^2}, \gamma x', y', z'\right)\right].$$

To make Q' look more like the expression for Q, change integration variables using $x = \gamma x'$, $y = y'$, and $z = z'$. Write the resulting expression for Q'. Confirm that for $v = 0$ you get $Q' = Q$.

(b) The strategy is to show that Q' is independent of the velocity v. In that case, the equality of Q' and Q for $v = 0$ implies the equality of Q' and Q for all velocities. Show that the derivative of Q' with respect to the velocity can be written as

$$\frac{dQ'}{dv} = \frac{1}{c^2} \int d^3x \left[x \left(\frac{\partial \rho}{\partial t} - \frac{v}{c^2} \frac{\partial j^x}{\partial t} \right) - j^x \right] \Bigg|_{t = \frac{vx}{c^2}, x, y, z}. \tag{1}$$

(c) Use the conservation equation $\partial_\alpha j^\alpha = 0$ to show that the integrand in (1) is a total derivative. It then follows from our conditions at infinity that $\frac{dQ'}{dv} = 0$, completing the proof that $Q' = Q$. You may find the following equation useful:

$$\left(\frac{\partial}{\partial x} j^x \left(\frac{vx}{c^2}, x, y, z \right) \right)_{y,z} = \left[\frac{v}{c^2} \frac{\partial j^x}{\partial t} + \frac{\partial j^x}{\partial x} \right] \Bigg|_{t = \frac{vx}{c^2}, x, y, z}, \tag{2}$$

where the subscripts y, z on the left-hand side indicate that the partial derivative is taken with these variables (and not t) held fixed. Equation (2) is just a statement about derivatives and should be clear if you understand the notation.

Problem 8.3 Angular momentum as a conserved charge.

Consider a Lagrangian L that depends only on the *magnitude* of the velocity $\dot{\vec{q}}(t)$ of a particle which moves in ordinary three-dimensional space.

(a) Write an infinitesimal variation $\delta\vec{q}(t)$ that represents a small rotation of the vector $\vec{q}(t)$. Explain why it leaves the Lagrangian invariant.
(b) Construct the conserved charge associated with this symmetry transformation. Verify that this conserved quantity is the (vector) angular momentum.

Problem 8.4 A generalization for charges and a special case for currents.

(a) Generalize equations (8.9) and (8.11) to the case in which there are various symmetry parameters ϵ^i and a set of coordinates q^a. Verify that your charges are conserved.
(b) Consider the setup of (8.17) for a "world" with no spatial dimensions. What are the possible values of the index α? What do equations (8.22), (8.23), and (8.25) give? Compare with your results in part (a).

Problem 8.5 Lorentz charges for the relativistic point particle.[†]

Consider the point particle action (5.15): $S = -mc\int d\tau\sqrt{-\eta_{\mu\nu}\dot{x}^\mu\dot{x}^\nu}$, where $\dot{x}^\mu(\tau) = dx^\mu(\tau)/d\tau$. This action can be treated as a mechanics action where x^μ is a coordinate and τ plays the role of time, or as a field action where x^μ is a field in a world with no spatial dimensions and with $\xi^0 = \tau$.

(a) Show that $x^\mu(\tau) \to x^\mu(\tau) + \epsilon^\mu$, with ϵ^μ constant, is a symmetry. Find the conserved charges associated with this symmetry and verify explicitly their conservation. Compare with the momenta $p_\mu(\tau)$ defined canonically from the Lagrangian.
(b) Write the infinitesimal Lorentz transformations of $x^\mu(\tau)$, and explain why the action is invariant under such transformations.
(c) Find an expression for the Lorentz charges in terms of $x^\mu(\tau)$ and $p^\mu(\tau)$. Make sure that your Lorentz charges coincide with angular momentum charges for the appropriate values of the indices. Verify explicitly their conservation.

Problem 8.6 Simple estimates regarding α', T_0, and ℓ_s.

(a) In hadronic physics $\alpha' \simeq 0.95\ \text{GeV}^{-2}$. Calculate the hadronic string tension in tons and the string length in centimeters.
(b) Assume that the string length is $\ell_s \simeq 10^{-30}$ cm. Calculate α' in GeV^{-2} and the string tension in tons.

Problem 8.7 Angular momentum of a rotating string.

Equation (8.51) tells us that the momentum $d\vec{p}$ carried by a small piece of string is given by

$$d\vec{p} = \frac{T_0 ds}{c^2}\,\frac{\vec{v}_\perp}{\sqrt{1-\frac{v_\perp^2}{c^2}}}$$

Use this result to calculate by direct integration the angular momentum J carried by a rotating open string. Make sure that your result coincides with (8.75).

Problem 8.8 Angular momentum of Kasey's jumping rope.

Consider the relativistic jumping rope solution of Problem 7.4. Calculate the magnitude J_z of the z-component of angular momentum (measured with respect to the origin) as a function of the separation L_0 and the angle γ. Show that $J_z/\hbar = (\sin^2\gamma)\alpha' E^2$, where E is the energy of the string.

Problem 8.9 Generalizing the construction of conserved charges.

We reconsider the setup that led to the conserved charge Q defined in (8.11). Assume now that the transformation (8.8) does not leave the Lagrangian L invariant, but rather the change is a total time derivative:

$$\delta L = \frac{d}{dt}(\epsilon\,\Lambda), \tag{1}$$

where Λ is some calculable function of coordinates, velocities, and possibly of time. Show that there is a modified conserved charge which takes the form

$$\epsilon\,Q = \frac{\partial L}{\partial \dot{q}}\,\delta q - \epsilon\,\Lambda. \tag{2}$$

Since it leads to a conserved charge, a transformation that changes L by a total derivative is said to be a symmetry transformation.

As an application consider a Lagrangian $L(q(t), \dot{q}(t))$ that has no explicit time dependence. A transformation of the form

$$q(t) \longrightarrow q(t+\epsilon) \simeq q(t) + \epsilon\dot{q}(t), \tag{3}$$

represents the effect of a constant infinitesimal time translation. Show that the transformation (3) is a symmetry in the sense of (1). Calculate Λ and construct the conserved charge Q. Is the result familiar?

Problem 8.10 Generalizing the construction of conserved currents.

Assume that the transformation (8.19) does not leave the Lagrangian density $\mathcal{L}$ invariant, but rather the change is a total derivative:

$$\delta\mathcal{L} = \frac{\partial}{\partial\xi^\alpha}(\epsilon^i\,\Lambda_i^\alpha), \tag{1}$$

where the Λ_i^α are a set of calculable functions of fields, field derivatives, and possibly coordinates. Show that there is a conserved current which takes the form

$$\epsilon^i\,j_i^\alpha = \frac{\partial\mathcal{L}}{\partial(\partial_\alpha\phi^a)}\delta\phi^a - \epsilon^i\,\Lambda_i^\alpha. \tag{2}$$

Since it leads to a conserved current, a transformation that changes $\mathcal{L}$ by a total derivative is said to be a symmetry transformation.

As an application consider a Lagrangian density $\mathcal{L}(\phi^a, \partial_\alpha \phi^a)$ that has no explicit dependence on the world coordinates ξ^α. The transformation

$$\phi^a(\xi^\beta) \longrightarrow \phi^a(\xi^\beta + \epsilon^\beta) \simeq \phi^a + \epsilon^\beta \partial_\beta \phi^a, \tag{3}$$

where ϵ^β is an infinitesimal constant vector, represents the effect of a constant translation. Show that (3) is a symmetry in the sense of (1), with

$$\Lambda^\alpha_\beta = \delta^\alpha_\beta \mathcal{L}. \tag{4}$$

Show that the conserved currents j^α_β take the form

$$j^\alpha_\beta = \frac{\partial \mathcal{L}}{\partial(\partial_\alpha \phi^a)} \partial_\beta \phi^a - \delta^\alpha_\beta \mathcal{L}. \tag{5}$$

Is j^0_0 familiar? The quantities j^α_β actually define the energy-momentum tensor T^α_β.

9 Light-cone relativistic strings

A class of gauges is introduced that fix the parameterization of the world-sheet, lead to a pair of constraints, and give equations of motion that are wave equations. One of these gauges, the light-cone gauge, sets X^+ proportional to τ and allows for a complete and explicit solution to the equations of motion. In this gauge, the dynamics of the string is determined by the motion along the transverse directions and the values of two zero modes. We encounter the classical Virasoro modes as the oscillation modes of the X^- coordinate, and we learn how to calculate the mass of an arbitrary string configuration.

9.1 A class of choices for τ

Our first encounter with classical string dynamics was simplified by our use of the static gauge. In this gauge the world-sheet time τ is identified with the spacetime time coordinate X^0 by

$$X^0(\tau, \sigma) = c\tau. \tag{9.1}$$

We are now going to consider more general gauges. Among the class of gauges that we will examine, one of them, the light-cone gauge, will turn out to be particularly useful. Using this gauge we will solve the equations of motion of the string in a complete and explicit fashion. Our solution in the static gauge was not fully explicit; the motion was characterized by a constrained vector function (see (7.48)).

The gauges we will focus on are those for which τ is set equal to a linear combination of the string coordinates. This condition can be written as

$$n_\mu \, X^\mu(\tau, \sigma) = \lambda\tau. \tag{9.2}$$

If we choose $n_\mu = (1, 0, \ldots, 0)$ and $\lambda = c$, this equation becomes (9.1). To understand the meaning of (9.2), we introduce the related equation

$$n_\mu \, x^\mu = \lambda\tau, \tag{9.3}$$

where we write x^μ, as opposed to X^μ, to emphasize that we are dealing with the general spacetime coordinates. Consider now the two equations above with the same *fixed* value of τ. If x_1^μ and x_2^μ are two points which satisfy (9.3), then $n_\mu(x_1^\mu - x_2^\mu) = 0$. This shows

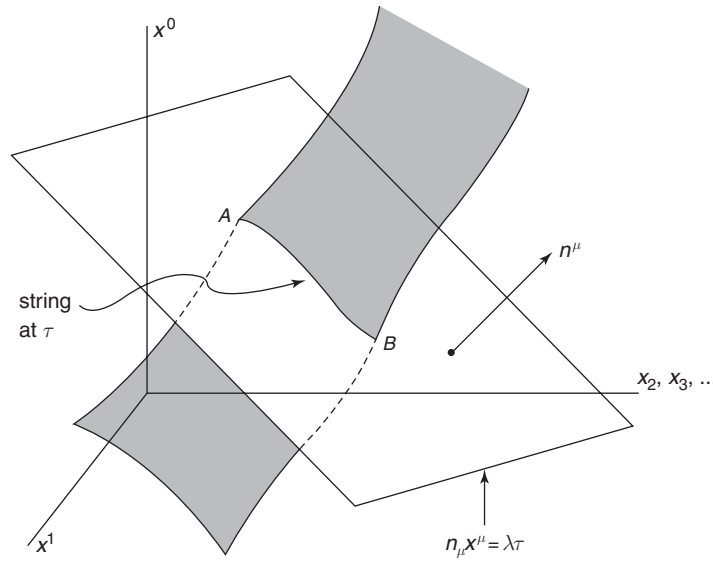

Fig. 9.1 **The gauge condition $n \cdot X = \lambda\tau$ fixes the strings to be the curves at the intersection of the world-sheet with the hyperplanes orthogonal to the vector n^μ.**

that any vector that joins points in the space (9.3) is orthogonal to the vector n^μ. The set of all points which satisfy (9.3) forms a hyperplane normal to n^μ.

We can now make clear the meaning of equation (9.2). The points X^μ that satisfy $n_\mu X^\mu = \lambda\tau$ are points that lie *both* on the world-sheet and on the hyperplane (9.3). Equation (9.2) states that all these points must be assigned the same value τ of the world-sheet time. If we define a string to be the set of points $X^\mu(\tau, \sigma)$ with constant τ, in the gauge (9.2), strings lie on hyperplanes of the form (9.3). The string with world-sheet time τ is the intersection of the world-sheet with the hyperplane $n \cdot x = \lambda\tau$, as illustrated in Figure 9.1.

We want strings to be spacelike objects. More precisely, the interval ΔX^μ between any two points on a string should be spacelike, perhaps null in some limit, but certainly never timelike. We will now see that in the gauge (9.2), a timelike n^μ guarantees that the string is spacelike. In this gauge any interval ΔX^μ along the string satisfies $n \cdot \Delta X = 0$. Since this condition is Lorentz invariant, it can be analyzed in a Lorentz frame where the only nonzero component of n^μ is the time component. In this frame ΔX^μ cannot have a time component. It is therefore a spacelike vector.

If n^μ is a null vector ($(1, 1, 0, \ldots, 0)$, for example), one can show that $n \cdot \Delta X = 0$ implies that ΔX^μ is generally spacelike and occassionally null (Problem 9.1). We will allow n^μ to be null in (9.2). This choice can be viewed as the limit of a sequence of choices all of which involve a timelike n^μ.

The gauges defined by (9.2) are not Lorentz covariant gauges for any choice of n^μ. Choosing n^μ selects a particular linear combination of spacetime coordinates to be set equal to τ. There is no linear combination of coordinates that is left invariant by arbitrary

Lorentz transformations. Therefore, the gauge condition takes different forms in different Lorentz frames – the gauge is not Lorentz covariant.

At this point, it is convenient to streamline the way that we deal with units. While we have been using the convention that τ and σ have units of time and length, respectively, starting now τ and σ will be *dimensionless*. We proved in Chapter 8 that for open strings with free endpoints, there is a well defined conserved momentum p^μ. We will use this Lorentz vector to rewrite our gauge condition (9.2) as follows:

$$n \cdot X(\tau, \sigma) = \tilde{\lambda}\,(n \cdot p)\,\tau. \tag{9.4}$$

Since $n \cdot p$ is a constant, the net effect is that we have traded λ for another constant $\tilde{\lambda}$. When open strings are attached to D-branes not all components of the string momentum are conserved. Since we want our analysis to hold even in this case, we will assume that the vector n^μ is chosen in such a way that $n \cdot p$ *is* conserved. This condition is weaker than the condition of momentum conservation. We will assume that $n \cdot \mathcal{P}^\sigma = 0$ at the open string endpoints since this condition naturally guarantees the conservation of $n \cdot p$ (consider equation (8.38) dotted with n^μ).

By involving the vector n^μ explicitly on both sides of equation (9.4), the scale of n^μ has been made irrelevant. Only the direction of n^μ matters. Bearing in mind that $n \cdot X$ has units of length, $n \cdot p$ has units of momentum, and τ is dimensionless, imagine dividing both sides of this equation by the unit of time. We then see that $\tilde{\lambda}$ has units of velocity divided by force. The canonical choice for velocity is c, and the canonical choice for force is the string tension T_0. It is therefore natural to set

$$\tilde{\lambda} \sim \frac{c}{T_0} = 2\pi\alpha'\hbar c^2, \tag{9.5}$$

where we used (8.76) to relate the string tension to α'. Before fixing $\tilde{\lambda}$ precisely, let us further simplify our treatment of units.

At stake is our ability to track the units of different physical quantities. We can simplify this matter by deciding to track just *one* unit, instead of the three units of length, time, and mass. By convention, we do this by setting two of the fundamental constants equal to one:

$$\hbar = c = 1, \tag{9.6}$$

as if these constants had no units! This has two consequences. First, any $\hbar$ or c in our formulae disappears without leaving a trace. This is not a serious problem, if we know the full units of an expression where $\hbar$ and c have been set equal to one, we can reconstruct the dependence on $\hbar$ and c unambiguously. Second, the units become dependent, and we are left with just one independent unit. Since $[c] = L/T$, the condition $c = 1$ implies that

$$L = T. \tag{9.7}$$

At this stage $[\hbar] = ML^2/T$ becomes $[\hbar] = ML$. With $\hbar = 1$ we get

$$M = 1/L\,. \tag{9.8}$$

Thus we can write all units in terms of mass or length (nobody uses time!). We will say that we are working with *natural units* when we set $\hbar = c = 1$ and track just one unit.

Back to (9.5), the units of α' have now become

$$[\alpha'] = \frac{1}{[T_0]} = \frac{L}{M} = L^2. \tag{9.9}$$

While the complete units of α' are those of inverse energy-squared, we see that in natural units, α' has units of length-squared. This is in agreement with our result in (8.78). In natural units the string length is

$$\ell_s = \sqrt{\alpha'}. \tag{9.10}$$

For reference, in natural units the Nambu–Goto action (8.77) takes the form

$$S = -\frac{1}{2\pi\alpha'} \int_{\tau_i}^{\tau_f} d\tau \int_0^{\sigma_1} d\sigma \sqrt{(\dot{X}\cdot X')^2 - (\dot{X})^2 (X')^2}\,. \tag{9.11}$$

In natural units equation (9.5) sets $\tilde{\lambda}$ proportional to α'. For open strings we choose $\tilde{\lambda} = 2\alpha'$, and equation (9.4) becomes

$$n \cdot X(\tau, \sigma) = 2\alpha'(n \cdot p)\,\tau \quad \text{(open strings)}. \tag{9.12}$$

This is the final form of the gauge condition which fixes the τ parameterization of the world-sheet.

When we use natural units, length scales can be expressed in terms of mass or energy scales. Given a length ℓ we can construct a unique mass m using only ℓ, $\hbar$, and c. This unique mass is $m = \hbar/(\ell c)$, or equivalently, $mc^2 = \hbar c/\ell$. For quick estimates we use $\hbar c \simeq 200\ \text{MeV} \times 10^{-15}\ \text{m}$.

Quick calculation 9.1 Show that the energy equivalent of the length 10^{-18} cm of a large extra dimension is roughly 20 TeV (1 TeV$= 10^{12}$ eV).

9.2 The associated σ parameterization

Having fixed the τ parameterization, let us determine the appropriate σ parameterization. In the static gauge the σ parameterization was fixed by the condition of constant energy density $\mathcal{P}^{\tau 0}$ over the strings (the curves of constant τ). Indeed, in setting $A(\sigma)$ in (7.17) equal to one by a suitable choice of σ, we made $\mathcal{P}^{\tau 0}$ in (7.6) constant. Since the static gauge uses $n_\mu = (1, 0, \ldots, 0)$, we were actually demanding the constancy of $n_\mu \mathcal{P}^{\tau\mu}$.

The proper generalization to situations where n^μ is arbitrary is to demand the constancy of $n_\mu \mathcal{P}^{\tau\mu} = n \cdot \mathcal{P}^\tau$ over the strings. Additionally, we require a parameterization range $\sigma \in [0, \pi]$ for all open strings. To show that we can satisfy these conditions we examine how $\mathcal{P}^{\tau\mu}(\tau, \sigma)$ transforms under σ reparameterizations. Looking at (6.93) we note that the rightmost fraction is invariant under σ reparameterizations, and only the derivative $\frac{ds}{d\sigma}$ transforms. It follows that given two parameterizations σ and $\tilde{\sigma}$ of a string we have

$$\mathcal{P}^{\tau\mu}(\tau, \sigma) = \frac{d\tilde{\sigma}}{d\sigma}\, \mathcal{P}^{\tau\mu}(\tau, \tilde{\sigma}) \quad \rightarrow \quad n \cdot \mathcal{P}^\tau(\tau, \sigma) = \frac{d\tilde{\sigma}}{d\sigma}\, n \cdot \mathcal{P}^\tau(\tau, \tilde{\sigma}). \tag{9.13}$$

If we are handed a string with $\tilde{\sigma}$ parameterization in which $n \cdot \mathcal{P}^\tau(\tau, \tilde{\sigma})$ depends on $\tilde{\sigma}$, we can choose a σ parameter so that $n \cdot \mathcal{P}^\tau(\tau, \sigma)$ does not depend on σ. This is simply done by adjusting the value of $\frac{d\tilde{\sigma}}{d\sigma}$ so that the final right-hand side above is set equal to a number that can only depend on τ. We then do a further reparameterization that scales σ by a constant factor b: $\sigma \rightarrow b\sigma$. This preserves the σ-independence of $n \cdot \mathcal{P}^\tau$ but allows us to have $\sigma \in [0, \pi]$, by adjusting b suitably. In this final parameterization we have

$$n \cdot \mathcal{P}^\tau(\tau, \sigma) = a(\tau), \tag{9.14}$$

where $a(\tau)$ is some function of τ. In fact $a(\tau)$ cannot depend on τ, and its value is already fixed by the conditions we have imposed. Integration of (9.14) over the string $\sigma \in [0, \pi]$ gives

$$\int_0^\pi d\sigma\, n \cdot \mathcal{P}^\tau(\tau, \sigma) = n \cdot p = \pi a(\tau) \quad \rightarrow \quad a(\tau) = \frac{n \cdot p}{\pi}. \tag{9.15}$$

Since $n \cdot p$ is conserved, $a(\tau)$ does not depend on τ. Back in (9.14) we have learned that

$$n \cdot \mathcal{P}^\tau = \frac{n \cdot p}{\pi} \quad \text{is an open string world-sheet constant.} \tag{9.16}$$

In this parameterization the momentum density $n \cdot \mathcal{P}^\tau$ is constant so the σ value assigned to a point is proportional to the amount of $n \cdot p$ momentum carried by the portion of string from the endpoint $\sigma = 0$ to the point.

Let us now examine the equation of motion $\partial_\tau \mathcal{P}^\tau_\mu + \partial_\sigma \mathcal{P}^\sigma_\mu = 0$. Dotting this equation with n^μ we arrive at the condition

$$\frac{\partial}{\partial\tau}(n \cdot \mathcal{P}^\tau) + \frac{\partial}{\partial\sigma}(n \cdot \mathcal{P}^\sigma) = 0. \tag{9.17}$$

The first term vanishes on account of (9.16), and we are left with

$$\frac{\partial}{\partial\sigma}(n \cdot \mathcal{P}^\sigma) = 0. \tag{9.18}$$

This means that $n \cdot \mathcal{P}^\sigma$ is independent of σ.

We now explain that for open strings, $n \cdot \mathcal{P}^\sigma = 0$. Because of equation (9.18), it suffices to show that $n \cdot \mathcal{P}^\sigma$ vanishes at some point on each string. As noted below equation (9.4), we assume $n \cdot \mathcal{P}^\sigma = 0$ at the string endpoints since it guarantees the conservation of $n \cdot p$. Thus, at least for open strings,

$$n \cdot \mathcal{P}^\sigma = 0. \tag{9.19}$$

Closed strings are dealt with in a similar way. We also parameterize them making the momentum density $n \cdot \mathcal{P}^\tau$ constant and, again, we conclude that $n \cdot \mathcal{P}^\tau$ is a *world-sheet* constant. For closed strings we want a range $\sigma \in [0, 2\pi]$, so (9.16) is changed into

$$n \cdot \mathcal{P}^\tau = \frac{n \cdot p}{2\pi} \quad \text{(closed strings)}. \tag{9.20}$$

Given this change, it will be convenient to write the τ parameterization of closed strings without the factor of two present in (9.12):

$$n \cdot X(\tau, \sigma) = \alpha'(n \cdot p)\,\tau \quad \text{(closed strings)}. \tag{9.21}$$

While equation (9.18) also holds for closed strings, we cannot prove that $n \cdot \mathcal{P}^\sigma = 0$; there is no special point on a closed string where $n \cdot \mathcal{P}^\sigma$ is known to vanish. We note, in addition, a related subtlety: it is not clear how to select the point $\sigma = 0$ at each value of τ. The two problems can be solved at once. Viewing the closed string world-sheet as a collection of closed strings (the curves of constant τ), we will select $\sigma = 0$ on *one* string arbitrarily. We will then select $\sigma = 0$ on all other strings by requiring $n \cdot \mathcal{P}^\sigma = 0$.

To show how this is done, we begin by using (6.50) to compute $n \cdot \mathcal{P}^\sigma$:

$$n \cdot \mathcal{P}^\sigma = -\frac{1}{2\pi\alpha'} \frac{(\dot{X} \cdot X')\partial_\tau(n \cdot X) - (\dot{X})^2 \partial_\sigma(n \cdot X)}{\sqrt{(\dot{X} \cdot X')^2 - (\dot{X})^2 (X')^2}}. \tag{9.22}$$

From (9.21) we see that $\partial_\sigma(n \cdot X) = 0$, and therefore $n \cdot \mathcal{P}^\sigma$ is

$$n \cdot \mathcal{P}^\sigma = -\frac{1}{2\pi\alpha'} \frac{(\dot{X} \cdot X')\,\partial_\tau(n \cdot X)}{\sqrt{(\dot{X} \cdot X')^2 - (\dot{X})^2 (X')^2}}. \tag{9.23}$$

It suffices to prove that we can make $n \cdot \mathcal{P}^\sigma$ vanish at one point on each string. Since $\partial_\tau(n \cdot X)$ is a constant, we must show that $\dot{X} \cdot X' = 0$ at some point on each string.

Pick an arbitrary point P on a given string, and declare it to be a point for which $\sigma = 0$ (Figure 9.2). We claim that there is (up to scaling) a unique vector v^μ tangent to the world-sheet at P and orthogonal to the spacelike vector $X^{\mu\prime}$ tangent to the string. This is not difficult to see. The world-sheet has a timelike tangent vector t^μ at P. Since $X^{\mu\prime}$ and t^μ are not parallel, they generate the tangent space to the world-sheet at P. If $t_\mu X^{\mu\prime} = 0$, then t^μ is the desired vector v^μ. If $t_\mu X^{\mu\prime} \neq 0$, we can define

$$v^\mu = t^\mu + bX^{\mu\prime}, \tag{9.24}$$

and solve for the constant b such that $v_\mu X^{\mu\prime} = 0$:

$$t \cdot X' + bX' \cdot X' = 0 \longrightarrow v^\mu = t^\mu - \frac{t \cdot X'}{X' \cdot X'} X^{\mu\prime}. \tag{9.25}$$

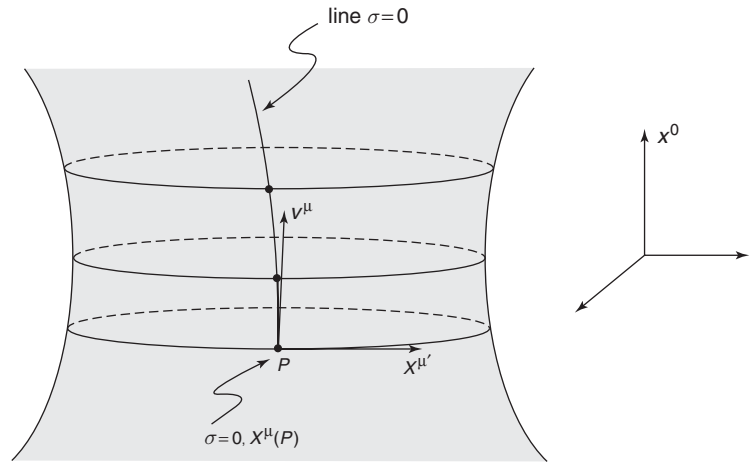

Fig. 9.2 **Fixing the $\sigma = 0$ line on the closed string world-sheet, after declaring P to be a point with $\sigma = 0$.**

Quick calculation 9.2 Show that the vector v^μ in (9.25) is timelike.

The point $\sigma = 0$ on neighboring strings is declared to be given by $X^\mu(P) + \epsilon\, v^\mu$, for infinitesimal ϵ. The vector v^μ is therefore tangent to the desired $\sigma = 0$ line at P. The full $\sigma = 0$ line is constructed by requiring that at each point its tangent be orthogonal to $X^{\mu\prime}$. Since the tangent to the $\sigma = 0$ line is proportional to $\dot{X}^\mu$, we guarantee that $\dot{X} \cdot X' = 0$ along the $\sigma = 0$ line. Thus we can ensure that $\dot{X} \cdot X'$, and consequently $n \cdot \mathcal{P}^\sigma$, vanishes at one point on each string. As a result, we have $n \cdot \mathcal{P}^\sigma = 0$ everywhere. Equation (9.19) can therefore be used both for open and for closed strings:

$$n \cdot \mathcal{P}^\sigma = 0 \quad \text{(open and closed strings).} \tag{9.26}$$

We have now concluded the description of the parameterizations of open and closed strings. The defining equations, in both cases, are summarized by

$$\begin{aligned} n \cdot X(\tau, \sigma) &= \beta \alpha' (n \cdot p)\, \tau, \\ n \cdot p &= \frac{2\pi}{\beta}\, n \cdot \mathcal{P}^\tau, \end{aligned} \tag{9.27}$$

where

$$\beta = \begin{cases} 2 & \text{for open strings,} \\ 1 & \text{for closed strings.} \end{cases} \tag{9.28}$$

Although we were able to define the $\sigma = 0$ line on the closed string world-sheet consistently, our construction has an obvious ambiguity. We had to choose one arbitrary point on one string. Any other point on that one string could have been used as $\sigma = 0$. This means that the parameterization of the closed string world-sheet can be shifted rigidly along the σ direction. There is no way to avoid this ambiguity. The gauge conditions do not fix uniquely the parameterization of the closed string world-sheet. This will have implications for the theory of closed strings.

9.3 Constraints and wave equations

Let us now explore the constraints on X' and $\dot{X}$ that are implied by our chosen parameterization. The vanishing of $n \cdot \mathcal{P}^\sigma$, together with (9.23) and the recognition that $\partial_\tau (n \cdot X)$ is a nonvanishing constant, leads to

$$\dot{X} \cdot X' = 0. \tag{9.29}$$

In the static gauge, $X^{0'} = 0$, and (9.29) reduces to $\dot{\vec{X}} \cdot \vec{X}' = 0$, which we obtained before in (7.1). Equation (9.29) is a constraint that follows from our parameterization.

We now use (9.29) to simplify the expression (6.49) for $\mathcal{P}^\tau$:

$$\mathcal{P}^{\tau\mu} = \frac{1}{2\pi\alpha'} \frac{X'^2 \dot{X}^\mu}{\sqrt{-\dot{X}^2 X'^2}}. \tag{9.30}$$

With the help of this result, the second equation in (9.27) gives

$$n \cdot p = \frac{1}{\beta\alpha'} \frac{X'^2 (n \cdot \dot{X})}{\sqrt{-\dot{X}^2 X'^2}}. \tag{9.31}$$

Since $n \cdot \dot{X} = \beta\alpha'(n \cdot p)$ (see (9.27)), the factors of β cancel, and we find

$$1 = \frac{X'^2}{\sqrt{-\dot{X}^2 X'^2}} \longrightarrow \dot{X}^2 + X'^2 = 0, \tag{9.32}$$

where we used $X'^2 \neq 0$. Aside from the units that are now different, this is consistent with the earlier result (7.23), which we obtained using the static gauge and involves only spatial components. Equations (9.29) and (9.32) are the constraint equations that follow from our choice of parameterization. Together they read

$$\dot{X} \cdot X' = 0, \quad \dot{X}^2 + X'^2 = 0. \tag{9.33}$$

These two conditions are conveniently packaged together as

$$(\dot{X} \pm X')^2 = 0. \tag{9.34}$$

Note that the constraints take this form for *any* value of β in (9.27). Our choices of β for open and closed strings give convenient σ ranges, but other choices are possible.

Given the above constraints, the momentum densities $\mathcal{P}^{\tau\mu}$ and $\mathcal{P}^{\sigma\mu}$ simplify considerably. To use the constraints to simplify (9.30), we must take the *positive* square root in the denominator. Since $X'^2 > 0$, using (9.32) gives

$$\sqrt{-\dot{X}^2 X'^2} = \sqrt{X'^2 X'^2} = X'^2. \tag{9.35}$$

Back in (9.30) we therefore have

$$\mathcal{P}^{\tau\mu} = \frac{1}{2\pi\alpha'}\dot{X}^\mu. \tag{9.36}$$

The momentum density $\mathcal{P}^{\sigma\mu}$, recorded in (6.50), simplifies down to

$$\mathcal{P}^{\sigma\mu} = \frac{1}{2\pi\alpha'}\frac{\dot{X}^2 X^{\mu\prime}}{\sqrt{-\dot{X}^2 X'^2}} = \frac{1}{2\pi\alpha'}\frac{\dot{X}^2 X^{\mu\prime}}{X'^2}, \tag{9.37}$$

and, using (9.32),

$$\mathcal{P}^{\sigma\mu} = -\frac{1}{2\pi\alpha'}X^{\mu\prime}. \tag{9.38}$$

The momentum densities are simple derivatives of the coordinates. Using these expressions in the field equation $\partial_\tau \mathcal{P}^{\tau\mu} + \partial_\sigma \mathcal{P}^{\sigma\mu} = 0$, we find

$$\ddot{X}^\mu - X^{\mu\prime\prime} = 0. \tag{9.39}$$

With our parameterization, the equations of motion are just wave equations! Notice that the minus sign on the right-hand side of (9.38) was necessary in order to get a wave equation. For open strings with free endpoints, the wave equations are supplemented by the requirement that the $\mathcal{P}^{\sigma\mu}$, and therefore the $X^{\mu\prime}$, vanish at the endpoints.

9.4 Wave equation and mode expansions

We will now explicitly solve the wave equation (9.39) in full generality for the case of open strings. In doing so we will introduce some of the basic notation that is used in string theory. We will assume that we have a space-filling D-brane. As a result, all string coordinates X^μ satisfy free boundary conditions at the endpoints. We know that the most general $X^\mu(\tau, \sigma)$ that solves the wave equation (9.39) is

$$X^\mu(\tau,\sigma) = \frac{1}{2}\Big(f^\mu(\tau+\sigma) + g^\mu(\tau-\sigma)\Big), \tag{9.40}$$

where f^μ and g^μ are arbitrary functions of a single argument. Bearing in mind (9.38), the free endpoint boundary conditions $\mathcal{P}^{\sigma\mu} = 0$ imply the Neumann boundary conditions

$$\frac{\partial X^\mu}{\partial\sigma} = 0, \quad \text{at} \quad \sigma = 0,\pi. \tag{9.41}$$

The boundary condition at $\sigma = 0$ gives us

$$\frac{\partial X^\mu}{\partial\sigma}(\tau,0) = \frac{1}{2}\Big(f^{\mu\prime}(\tau) - g^{\mu\prime}(\tau)\Big) = 0. \tag{9.42}$$

Since the derivatives of f^μ and g^μ coincide, f^μ and g^μ can differ only by a constant c^μ. After replacing $g^\mu = f^\mu + c^\mu$ in (9.40), the constant c^μ can be reabsorbed into the definition of f^μ. The result is

$$X^\mu(\tau,\sigma) = \frac{1}{2}\Big(f^\mu(\tau+\sigma) + f^\mu(\tau-\sigma)\Big). \tag{9.43}$$

Now let us consider the boundary condition at $\sigma = \pi$:

$$\frac{\partial X^\mu}{\partial\sigma}(\tau,\pi) = \frac{1}{2}\Big(f^{\mu\prime}(\tau+\pi) - f^{\mu\prime}(\tau-\pi)\Big) = 0. \tag{9.44}$$

Since this equation must hold for all τ, we learn that *$f^{\mu\prime}$ is periodic with period 2π*. Since 2π is a natural period, our decision to parameterize the open string with $\sigma \in [0,\pi]$ has paid off.

We now write the general Fourier series for the periodic function $f^{\mu\prime}(u)$:

$$f^{\mu\prime}(u) = f_1^\mu + \sum_{n=1}^{\infty}\Big(a_n^\mu \cos nu + b_n^\mu \sin nu\Big). \tag{9.45}$$

Integrating this equation we get the expansion of $f^\mu(u)$:

$$f^\mu(u) = f_0^\mu + f_1^\mu u + \sum_{n=1}^{\infty}\Big(A_n^\mu \cos nu + B_n^\mu \sin nu\Big), \tag{9.46}$$

where we have absorbed the constants arising from integration into new coefficients. We substitute this expression for $f(u)$ back into (9.43) and simplify to get

$$X^\mu(\tau,\sigma) = f_0^\mu + f_1^\mu\tau + \sum_{n=1}^{\infty}\Big(A_n^\mu \cos n\tau + B_n^\mu \sin n\tau\Big)\cos n\sigma. \tag{9.47}$$

We want to replace the coefficients in equation (9.47) by new coefficients that have a simple physical interpretation. Our first step is introducing constants a_n^μ through the relations

$$\begin{aligned} A_n^\mu \cos n\tau + B_n^\mu \sin n\tau &= -\frac{i}{2}\left((B_n^\mu + iA_n^\mu)e^{in\tau} - (B_n^\mu - iA_n^\mu)e^{-in\tau}\right) \\ &\equiv -i\,\frac{\sqrt{2\alpha'}}{\sqrt{n}}\left(a_n^{\mu *}e^{in\tau} - a_n^\mu e^{-in\tau}\right). \end{aligned} \tag{9.48}$$

Here $*$ denotes complex conjugation. The purpose of the $\sqrt{2\alpha'}$ factor is to make the constants a_n^μ dimensionless. These constants, and their complex conjugates, will turn into annihilation and creation operators when we consider the quantum theory. Equation (9.48) introduces the notation that string theorists conventionally use.

In equation (9.47) the constant f_1^μ has a simple physical interpretation. Using (9.36), the momentum density is given by

$$\mathcal{P}^{\tau\mu} = \frac{1}{2\pi\alpha'}\,\dot{X}^\mu = \frac{1}{2\pi\alpha'}\,f_1^\mu + \cdots, \tag{9.49}$$

where the dots denote terms with $\cos n\sigma$ dependence ($n \neq 0$). To find the total momentum p^μ, we integrate $\mathcal{P}^{\tau\mu}$ over $\sigma \in [\,0, \pi\,]$. Happily, the terms represented by the dots do not contribute as the integral of $\cos n\sigma$ vanishes. We get

$$p^\mu = \int_0^\pi \mathcal{P}^{\tau\mu} d\sigma = \frac{1}{2\pi\alpha'}\,\pi f_1^\mu \;\longrightarrow\; f_1^\mu = 2\alpha' p^\mu. \tag{9.50}$$

This identifies f_1^μ as a quantity proportional to the spacetime momentum carried by the string. Declaring $f_0^\mu = x_0^\mu$, and collecting all the above information, equation (9.47) now takes the conventional form

$$X^\mu(\tau,\sigma) = x_0^\mu + 2\alpha' p^\mu \tau - i\sqrt{2\alpha'}\sum_{n=1}^{\infty}\Big(a_n^{\mu *}\, e^{in\tau} - a_n^\mu e^{-in\tau}\Big)\frac{\cos n\sigma}{\sqrt{n}}. \tag{9.51}$$

The terms on the right-hand side clearly correspond to the zero mode, to the momentum, and to the oscillations of the string. If all the coefficients a_n^μ of the oscillations vanish, the equation represents the motion of a point particle.

Quick calculation 9.3 Verify explicitly that $X^\mu(\tau,\sigma)$ is real.

Let us now introduce some notation that will allow us to write simple expressions for the τ and σ derivatives of $X^\mu(\tau,\sigma)$. We start by defining

$$\alpha_0^\mu = \sqrt{2\alpha'}p^\mu. \tag{9.52}$$

Furthermore, we define

$$\alpha_n^\mu = a_n^\mu\sqrt{n}, \quad \alpha_{-n}^\mu = a_n^{\mu *}\sqrt{n}, \quad n \geq 1. \tag{9.53}$$

It is important to note that

$$\alpha_{-n}^\mu = (\alpha_n^\mu)^*. \tag{9.54}$$

Moreover, while the a_n^μ are only defined when n is a positive integer, the α_n^μ are defined for any integer n, including zero. Using these new names, we can rewrite X^μ as

$$X^\mu(\tau,\sigma) = x_0^\mu + \sqrt{2\alpha'}\,\alpha_0^\mu \tau - i\sqrt{2\alpha'}\sum_{n=1}^{\infty}\frac{1}{n}\left(\alpha_{-n}^\mu e^{in\tau} - \alpha_n^\mu e^{-in\tau}\right)\cos n\sigma. \tag{9.55}$$

It is convenient to sum over all integers except zero:

$$X^\mu(\tau,\sigma) = x_0^\mu + \sqrt{2\alpha'}\,\alpha_0^\mu \tau + i\sqrt{2\alpha'}\sum_{n\neq 0}\frac{1}{n}\,\alpha_n^\mu\, e^{-in\tau}\cos n\sigma. \tag{9.56}$$

This completes the solution of the wave equations with Neumann boundary conditions. In the above equation, a solution is defined once we specify the constants x_0^μ and α_n^μ for $n \geq 0$.

It is convenient to record here the τ and σ derivatives of X^μ. From (9.56) we see that

$$\dot{X}^\mu = \sqrt{2\alpha'}\sum_{n\in\mathbb{Z}}\alpha_n^\mu \cos n\sigma e^{-in\tau}, \tag{9.57}$$

$$X^{\mu\prime} = -i\sqrt{2\alpha'}\sum_{n\in\mathbb{Z}}\alpha_n^\mu \sin n\sigma e^{-in\tau}, \tag{9.58}$$

where $\mathbb{Z}$ denotes the set of all integers (positive, negative, and zero). Finally, two linear combinations of the above derivatives are particularly nice:

$$\dot{X}^\mu \pm X^{\mu\prime} = \sqrt{2\alpha'}\sum_{n\in\mathbb{Z}}\alpha_n^\mu\, e^{-in(\tau\pm\sigma)}. \tag{9.59}$$

We have found solutions of the wave equations that satisfy the relevant boundary conditions, but we must also ensure that the constraints (9.33) are satisfied. If we specify arbitrarily all constants α_n^μ, the constraints will not be satisfied. We will use the light-cone gauge to find a solution that satisfies the wave equations as well as the constraints.

9.5 Light-cone solution of equations of motion

The light-cone solution of the equations of motion involves using light-cone coordinates to represent the motion of strings, and imposing a set of conditions that defines the light-cone gauge. We have seen in Chapter 2 that using light-cone coordinates means using x^+ and x^- instead of x^0 and x^1 – this is just a change of coordinates. Imposing a light-cone gauge condition is a more substantial step. The gauges we have examined in this chapter

represent very specific choices of world-sheet coordinates. One of these choices is the light-cone gauge.

Selecting the light-cone gauge means imposing the conditions (9.27) with a vector n^μ that gives $n \cdot X = X^+$. Taking

$$n_\mu = \Big(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, \ldots, 0\Big), \tag{9.60}$$

we indeed find

$$n \cdot X = \frac{X^0 + X^1}{\sqrt{2}} = X^+, \quad n \cdot p = \frac{p^0 + p^1}{\sqrt{2}} = p^+. \tag{9.61}$$

Using these relations in (9.27) we have

$$X^+(\tau, \sigma) = \beta\alpha' p^+ \tau, \quad p^+ = \frac{2\pi}{\beta}\, \mathcal{P}^{\tau +}, \tag{9.62}$$

where $\beta = 2$ for open strings and $\beta = 1$ for closed strings. The second equation tells us that the density of p^+ is constant along the string.

The strategy behind the light-cone gauge is to use the especially simple form of X^+ to show that there is no dynamics in X^- (up to a zero mode) and that all the dynamics is in the *transverse* coordinates $X^2, X^3, \ldots, X^d$. These transverse coordinates will be denoted by X^I, where the transverse index I runs from 2 up to d:

$$X^I = (X^2, X^3, \ldots, X^d). \tag{9.63}$$

In order to proceed we look at the constraint equations (9.34). Using the definition (2.59) of the relativistic dot product in light-cone coordinates, we can write these constraints as

$$-2(\dot{X}^+ \pm X^{+\prime})(\dot{X}^- \pm X^{-\prime}) + (\dot{X}^I \pm X^{I\prime})^2 = 0, \tag{9.64}$$

where $(a^I)^2 = a^I a^I$ and, as usual, repeated indices imply summation. Since $X^{+\prime} = 0$ and $\dot{X}^+ = \beta\alpha' p^+$, we in fact have

$$\dot{X}^- \pm X^{-\prime} = \frac{1}{\beta\alpha'} \frac{1}{2p^+} (\dot{X}^I \pm X^{I\prime})^2. \tag{9.65}$$

In writing the above we have assumed that $p^+ \neq 0$. While p^+ certainly satisfies $p^+ \geq 0$, it can happen that p^+ is equal to zero. For this, the momentum p^1 must cancel the energy, and this can only occur if we have a massless particle traveling exactly in the negative x^1 direction. Since the vanishing of p^+ is thus not a common occurrence, we will take p^+ to be always positive. If we come across a situation where p^+ is zero, the light-cone formalism will not apply.

Note the crucial role played by both the choice of light-cone coordinates and the choice of light-cone gauge in allowing us to solve for the derivatives of X^-. Light-cone coordinates were useful because the off-diagonal metric in the $(+, -)$ sector allowed us to solve for the derivatives of X^- without having to take a square root! We just had to divide by $\dot{X}^+$. Here the light-cone *gauge* was useful, since it made $\dot{X}^+$ equal to a constant.

Equations (9.65) determine both $\dot{X}^-$ and $X^{-\prime}$ in terms of the X^I, and therefore they determine X^- up to a single integration constant. All that is needed is the value of X^- at some point P on the world-sheet and then we can integrate the relation

$$dX^- = \frac{\partial X^-}{\partial \tau}\, d\tau + \frac{\partial X^-}{\partial \sigma}\, d\sigma, \tag{9.66}$$

to find the value of X^- at any other point Q. On an open string world-sheet we can choose any path from P to Q to perform the integration and the result for $X^-(Q)$ will not depend on the path, as discussed in Problem 9.2. On a closed string world-sheet there is a further consistency condition. Imagine a contour of integration that starts at P and ends at P after going around the world-sheet. It is not guaranteed that the integration of dX^- on this contour gives zero, a result necessary for X^- to be well-defined. If we choose the contour to be at constant τ, we must require

$$\int_0^{2\pi} d\sigma\, \frac{\partial X^-}{\partial \sigma} = 0. \tag{9.67}$$

This is a nontrivial constraint; see Problem 9.5.

Our analysis indicates that the full evolution of the string is determined by the following set of objects:

$$X^I(\tau, \sigma), \quad p^+, \; x_0^-, \tag{9.68}$$

where x_0^- is the constant of integration needed for X^-.

Let us focus on the case of open strings ($\beta = 2$). We consider the explicit solution for the transverse coordinates X^I and calculate the associated X^-. Making use of the general solution in (9.56) we have

$$X^I(\tau, \sigma) = x_0^I + \sqrt{2\alpha'}\, \alpha_0^I \tau + i\sqrt{2\alpha'} \sum_{n \neq 0} \frac{1}{n} \alpha_n^I e^{-in\tau} \cos n\sigma. \tag{9.69}$$

Moreover, for the X^+ coordinate the gauge condition gives

$$X^+(\tau, \sigma) = 2\alpha' p^+ \tau = \sqrt{2\alpha'}\, \alpha_0^+ \tau. \tag{9.70}$$

As we can see, the position zero mode and the oscillations of the X^+ coordinate have been set to equal to zero:

$$x_0^+ = 0, \quad \alpha_n^+ = \alpha_{-n}^+ = 0, \quad n = 1, 2, \ldots, \infty. \tag{9.71}$$

What about X^-? Being a linear combination of X^0 and X^1, the coordinate X^- satisfies the same wave equation and the same boundary conditions as all the other coordinates. We can therefore use the same expansion as in (9.56) to write

$$X^-(\tau, \sigma) = x_0^- + \sqrt{2\alpha'}\, \alpha_0^- \tau + i\sqrt{2\alpha'} \sum_{n \neq 0} \frac{1}{n} \alpha_n^- e^{-in\tau} \cos n\sigma. \tag{9.72}$$

Using equation (9.59) with $\mu = -$ and $\mu = I$, we find

$$\dot{X}^- \pm X^{-\prime} = \sqrt{2\alpha'} \sum_{n\in\mathbb{Z}} \alpha_n^- \, e^{-in(\tau\pm\sigma)}, \tag{9.73}$$

$$\dot{X}^I \pm X^{I\prime} = \sqrt{2\alpha'} \sum_{n\in\mathbb{Z}} \alpha_n^I \, e^{-in(\tau\pm\sigma)}. \tag{9.74}$$

We use these equations and (9.65) to solve for the minus oscillators:

$$\begin{aligned}\sqrt{2\alpha'} \sum_{n\in\mathbb{Z}} \alpha_n^- \, e^{-in(\tau\pm\sigma)} &= \frac{1}{2p^+} \sum_{p,q\in\mathbb{Z}} \alpha_p^I \alpha_q^I e^{-i(p+q)(\tau\pm\sigma)} \\ &= \frac{1}{2p^+} \sum_{n,p\in\mathbb{Z}} \alpha_p^I \alpha_{n-p}^I e^{-in(\tau\pm\sigma)} \\ &= \frac{1}{2p^+} \sum_{n\in\mathbb{Z}} \Big(\sum_{p\in\mathbb{Z}} \alpha_p^I \alpha_{n-p}^I \Big) \, e^{-in(\tau\pm\sigma)}. \end{aligned} \tag{9.75}$$

It follows that we can identify α_n^- consistently as

$$\sqrt{2\alpha'}\, \alpha_n^- = \frac{1}{2p^+} \sum_{p\in\mathbb{Z}} \alpha_{n-p}^I \alpha_p^I. \tag{9.76}$$

This represents a complete solution! We now have explicit expressions for the minus oscillators α_n^- in terms of the transverse oscillators. On the right-hand side the spacetime indices are only to be summed over the labels of the transverse coordinates.

The general solution which represents an allowed motion is fixed by specifying arbitrary values for p^+, x_0^-, x_0^I, and for all the constants α_n^I. This clearly determines $X^I(\tau,\sigma)$ in (9.69), and $X^+(\tau,\sigma)$ in (9.70). Using (9.76) we can calculate the constants α_n^-, which together with x_0^- determine $X^-(\tau,\sigma)$ in (9.72). The full solution is thus constructed.

The quadratic combination of oscillators on the right-hand side of (9.76) is remarkably useful, so it has been given a name. It is the *transverse Virasoro* mode $L_n^\perp$:

$$\sqrt{2\alpha'}\, \alpha_n^- = \frac{1}{p^+} L_n^\perp, \quad L_n^\perp \equiv \frac{1}{2} \sum_{p\in\mathbb{Z}} \alpha_{n-p}^I \alpha_p^I. \tag{9.77}$$

In particular, for $n = 0$ we use (9.52) and find

$$\sqrt{2\alpha'}\, \alpha_0^- \;=\; 2\alpha' p^- \;=\; \frac{1}{p^+} L_0^\perp \;\longrightarrow\; \quad 2p^+ p^- = \frac{1}{\alpha'} L_0^\perp. \tag{9.78}$$

Using the value of α_n^- given in (9.77), equations (9.73) and (9.65) are written as

$$\dot{X}^- \pm X^{-\prime} = \frac{1}{p^+}\sum_{n\in\mathbb{Z}} L_n^\perp\, e^{-in(\tau\pm\sigma)} = \frac{1}{4\alpha' p^+}\,(\dot{X}^I \pm X^{I\prime})^2. \tag{9.79}$$

Quick calculation 9.4 Show that

$$X^-(\tau,\sigma) = x_0^- + \frac{1}{p^+}L_0^\perp\tau + \frac{i}{p^+}\sum_{n\neq 0}\frac{1}{n}L_n^\perp e^{-in\tau}\cos n\sigma. \tag{9.80}$$

This equation explicitly demonstrates the claim that the Virasoro modes are the expansion modes of the coordinate $X^-(\tau,\sigma)$.

It is instructive to compute the mass of a string which is performing an arbitrary motion. The mass can be calculated using the relativistic equation

$$M^2 = -p^2 = 2p^+p^- - p^Ip^I. \tag{9.81}$$

Since the mass is a constant of the motion, we anticipate that it depends on the constant coefficients a_n^I introduced to define a classical solution. To evaluate the mass we start with (9.78), substitute the value of $L_0^\perp$ from (9.77), and use the definitions in (9.52) and (9.53):

$$2p^+p^- = \frac{1}{\alpha'}L_0^\perp = \frac{1}{\alpha'}\left(\frac{1}{2}\alpha_0^I\alpha_0^I + \sum_{n=1}^{\infty}\alpha_n^{I*}\,\alpha_n^I\right) = p^Ip^I + \frac{1}{\alpha'}\sum_{n=1}^{\infty} n\,a_n^{I*}\,a_n^I. \tag{9.82}$$

Replacing this result on the right-hand side of (9.81) we finally find

$$M^2 = \frac{1}{\alpha'}\sum_{n=1}^{\infty} n\,a_n^{I*}\,a_n^I. \tag{9.83}$$

This is a very interesting result. The mass-squared is written as a sum of terms each of which is of the form $a^*a = |a|^2 \geq 0$. So, we find that $M^2 \geq 0$. This shows that the classical string mass $M = \sqrt{M^2}$ is a real number (conventionally taken to be positive). Such a result is actually hard to obtain without using the light-cone gauge. We also see that we can adjust the coefficients a_n^I to obtain classical string solutions with arbitrary values of the mass. If all the coefficients a_n^I vanish, the result is a massless object $M^2 = 0$. Indeed, when all a_n^I vanish the string collapses to a moving point: equation (9.69) gives $X^I(\tau,\sigma) = x_0^I + \sqrt{2\alpha'}\,\alpha_0^I\tau$, and the σ dependence disappears.

Quick calculation 9.5 Calculate $X^-(\tau,\sigma)$ when all a_n^I vanish. Note that the σ dependence of X^- disappears.

The classical result (9.83) for M^2 does not survive quantization. First, M^2 will become quantized, and string states will not exhibit a continuous spectrum of masses. This is good

because we do not observe in nature particle states that take continuous values of the mass. Even more, equation (9.83) does not give enough massless states. The few massless states that are obtained do not behave at all like the massless states of Maxwell theory. For closed strings, the few massless states that are obtained by a similar analysis do not behave at all like the massless states of gravitation. Quantum mechanics will add an extra constant to the formula for M^2, both for open and for closed strings. These additive constants will enable us to find states that correspond to those of physical theories. String theory has a chance to describe gauge fields and gravity because quantization changes (9.83) and the corresponding formula for closed strings.

Problems

Problem 9.1 Vectors orthogonal to null vectors are null or spacelike.

Let n^μ be a nonzero null vector ($n_\mu n^\mu = 0$) in D-dimensional Minkowski space. In addition, let b^μ be a vector that satisfies $n_\mu b^\mu = 0$. Prove the following.

(a) The vector b^μ is either spacelike or null.
(b) If b^μ is null, then $b^\mu = \lambda n^\mu$ for some constant λ.
(c) The set of vectors b^μ that satisfies $n_\mu b^\mu = 0$ is a vector space V of dimension $(D-1)$. The subset of null vectors b^μ is a vector subspace of V of dimension one.

This result shows that for gauges (9.2) with n^μ null and for $D > 2$, strings are almost always spacelike objects. Moreover, the hyperplane orthogonal to n^μ contains n^μ. This is readily confirmed in two dimensions:

(d) Let $D = 2$ and consider a spacetime diagram such as the one in Figure 2.2. What is the null vector n^μ such that $n \cdot X = X^+$? Confirm that n^μ points along the lines of constant X^+.

Problem 9.2 Consistency checks on the solution for X^-.

(a) Use (9.65) to find $\partial_\tau X^-$ and $\partial_\sigma X^-$. Show that the consistency condition $\partial_\sigma(\partial_\tau X^-) = \partial_\tau(\partial_\sigma X^-)$ holds if the transverse coordinates X^I satisfy the wave equation. Prove that this condition guarantees that X^-, determined by integration of (9.66), is independent of the chosen path.
(b) Show that X^-, as calculated in (9.65), satisfies the wave equation if the transverse coordinates X^I satisfy the wave equation.
(c) Assume that at the open string endpoints some of the transverse light-cone coordinates X^I satisfy Neumann boundary conditions and some satisfy Dirichlet boundary conditions. Prove that X^-, as calculated in (9.65), will always satisfy Neumann boundary conditions.

Problem 9.3 Rotating open string in the light-cone gauge.

Consider string motion defined by $x_0^- = x_0^I = 0$, and the vanishing of all coefficients α_n^I with the exception of

$$\alpha_1^{(2)} = \alpha_{-1}^{(2)*} = a, \quad \alpha_1^{(3)} = \alpha_{-1}^{(3)*} = ia. \tag{1}$$

Here a is a dimensionless real constant. We want to construct a solution that represents an open string that is rotating on the (x^2, x^3) plane.

(a) What is the mass (or energy) of this string?
(b) Construct the explicit functions $X^2(\tau, \sigma)$ and $X^3(\tau, \sigma)$. What is the length of the string in terms of a and α'?
(c) Calculate the $L_n^\perp$ modes for all n. Use your result to construct $X^-(\tau, \sigma)$. Your answer should be σ-independent!
(d) Determine the value of p^+ using the condition that for this string $X^1(\tau, \sigma) = 0$. Find the relation between t and τ.
(e) Confirm that in your solution the energy of the string and its angular frequency of rotation are related to its length as in (7.59).

Problem 9.4 Generating consistent open string motion: How does an open string collapse?

Consider string motion defined by $x_0^- = x_0^I = 0$, and the vanishing of all coefficients α_n^I with the exception of

$$\alpha_1^{(2)} = \alpha_{-1}^{(2)*} = a.$$

Here a is a dimensionless real constant. We want to construct a solution that represents an open string oscillating on the (x^1, x^2) plane and having *zero* momentum in this plane.

(a) Show that the string motion is described by

$$\begin{aligned}
\frac{1}{\sqrt{2\alpha'}}\frac{1}{a}X^0(\tau, \sigma) &= \sqrt{2}\left(\tau + \frac{1}{4}\sin 2\tau \cos 2\sigma\right), \\
\frac{1}{\sqrt{2\alpha'}}\frac{1}{a}X^1(\tau, \sigma) &= -\frac{1}{2\sqrt{2}}\sin 2\tau \cos 2\sigma, \\
\frac{1}{\sqrt{2\alpha'}}\frac{1}{a}X^2(\tau, \sigma) &= 2\sin\tau \cos\sigma.
\end{aligned}$$

(b) Confirm that τ flows as t flows. In the chosen Lorentz frame, strings are lines on the world-sheet that lie at constant time X^0. Find the values of τ for which constant τ lines are strings. Describe those strings.
(c) At $\tau = 0$ the string has zero length. Study in detail the motion for $\tau \ll 1$. Calculate $\tau = \tau(t, \sigma)$ and use this result to find $X^1(t, \sigma)$ and $X^2(t, \sigma)$. Prove that as the string expands from zero size, it lies on the portion $\cos\theta \geq -1/3$ of a circle centered at the origin, whose radius grows at the speed of light (θ is measured with respect to the positive x^1 axis). Note that the endpoints move transversely to the string.
(d) Use your favorite software package to do a parametric plot of the string world-sheet as a surface in three dimensions. Use X^1, X^2, and X^0 as x, y, and z axes, respectively, and parameters τ and σ. For further help visualizing the motion of the string, plot the string on the (x^1, x^2) plane at various values of the time X^0. This requires solving (numerically) for τ as a function of X^0 and σ.

Problem 9.5 A closed string in the light-cone gauge.

Consider a closed string for which

$$X^{(2)}(\tau,\sigma) = \sqrt{2\alpha'}\Big(a\sin(\tau-\sigma) + b\cos(\tau-\sigma) + \bar{a}\sin(\tau+\sigma) + \bar{b}\cos(\tau+\sigma)\Big),$$

and all other transverse coordinates $X^I(\tau,\sigma)$ vanish. In the above, a, b, $\bar{a}$, and $\bar{b}$ are real constants. Note that a and b are the coefficients of waves that propagate towards increasing σ while $\bar{a}$ and $\bar{b}$ are the coefficients of waves that propagate towards decreasing σ. As usual for closed strings, we set $X^+(\tau,\sigma) = \alpha' p^+\tau$ and use $\sigma \in [0, 2\pi]$.

(a) As it turns out, it is not possible to generate a solution of the equations of motion for completely arbitrary values of the constants a, b, $\bar{a}$, and $\bar{b}$. Examine the calculation of $X^-(\tau,\sigma)$ and derive the constraint that the constants must satisfy.

(b) Your constraint must allow $a = b = \bar{a} = \bar{b} = r$. Use these values to calculate $X^-(\tau,\sigma)$ and to determine the mass of this string.

10 Light-cone fields and particles

We study the classical equations of motion for scalar fields, Maxwell fields, and gravitational fields. We use the light-cone gauge to find plane-wave solutions to their equations of motion and the number of degrees of freedom that characterize them. We explain how the quantization of such classical field configurations gives rise to particle states – scalar particles, photons, and gravitons. In doing so we prepare the ground for the later identification of such states among the quantum states of relativistic strings.

10.1 Introduction

In our investigation of classical string motion we had a great deal of freedom in choosing the coordinates on the world-sheet. This freedom was a direct consequence of the reparameterization invariance of the action, and we exploited it to simplify the equations of motion tremendously. Reparameterization invariance is an example of a *gauge invariance*, and a choice of parameterization is an example of a choice of *gauge*. We saw that the light-cone gauge – a particular parameterization in which τ is related to the light-cone time X^+ and σ is chosen so that the p^+-density is constant – was useful to obtain a complete and explicit solution of the equations of motion.

Classical field theories sometimes have gauge invariances. Classical electrodynamics, for example, is described in terms of gauge potentials A_μ. The gauge invariance of this description is often used to great advantage. The classical theory of a *scalar* field is simpler than classical electromagnetism. This theory is not studied at the undergraduate level, however, because elementary scalar particles – the kind of particles associated with the quantum theory of scalar fields – have not been detected yet. On the other hand, photons – the particles associated with the quantum theory of the electromagnetic field – are found everywhere! Scalar particles may play an important role in the Standard Model of particle physics, where they can help trigger symmetry breaking. Thus physicists may detect scalar particles some time in the future. The field theory of a single scalar field has no gauge invariance. We will study it because it is the simplest field theory and because scalar particles arise in string theory. The most famous scalar particle in string theory is the tachyon. Also important is the dilaton, a massless scalar particle.

Einstein's classical field theory of gravitation is more complicated than classical electromagnetism. In gravity, as we explained in Section 3.6, the dynamical variable is the

two-index metric field $g_{\mu\nu}(x)$. Gravitation has a very large gauge invariance. The gauge transformations involve reparameterizations of spacetime.

We will consider scalar fields, electromagnetic fields, and gravitational fields. The light-cone gauge will allow us to simplify dramatically the (linearized) equations of motion, find their plane-wave solutions, and count the number of degrees of freedom that characterize the solutions. We will also briefly consider how the quantization of plane-wave solutions gives rise to particle states. These are the quantum states associated with the field theories. We quantize the relativistic string in Chapter 12. There we relate the quantum states of the string to the particle states of the field theories that we study in the present chapter. We use the light-cone gauge here because we will quantize the relativistic string in the light-cone gauge.

10.2 An action for scalar fields

A scalar field is simply a single real function of spacetime. It is written as $\phi(t, \vec{x})$ or, more briefly, as $\phi(x)$. The term scalar means scalar under Lorentz transformations: all Lorentz observers will agree on the value of the scalar field at any fixed point in spacetime. Scalar fields have no Lorentz indices.

Let us now motivate the simplest kind of action principle that can be used to define the dynamics of a scalar field. Consider first the kinetic energy. In mechanics, the kinetic energy of a particle is proportional to its velocity squared. For a scalar field, the kinetic energy density T is declared to be proportional to the square of the rate of change of the field with time:

$$T = \frac{1}{2}(\partial_0\phi)^2 . \tag{10.1}$$

We speak of densities because, at any fixed time, T is a function of position. The total kinetic energy will be the integral of the density T over space.

Now consider the potential energy density. There is one class of term that is natural. Suppose the equilibrium value of the field is $\phi = 0$. For a simple harmonic oscillator with equilibrium position $x = 0$, the potential energy goes like $V \sim x^2$. If we want the field to prefer its equilibrium state, then this must be encoded in the potential. The simplest potential which does this is quadratic:

$$V = \frac{1}{2}m^2\phi^2 . \tag{10.2}$$

It is interesting to note that the constant m introduced here has the units of mass. Indeed, since the expressions on the right-hand side of the two equations above must have the same units ($[T] = [V]$), and both have two factors of ϕ, we require $[m] = [\partial_0] = L^{-1} = M$.

We could now attempt to form a Lagrangian density by combining the two energies above and setting

$$\mathcal{L} \stackrel{?}{=} T - V = \frac{1}{2}(\partial_0\phi)^2 - \frac{1}{2}m^2\phi^2 . \tag{10.3}$$

This Lagrangian density, however, is not Lorentz invariant. The second term on the right-hand side is a scalar, but the first term is not, for it treats time as special. We are missing a contribution representing the energy cost when the scalar field varies in space. This is eminently reasonable in special relativity: if it costs energy to have the field vary in time, it must also cost energy to have the field vary in space. The extra contribution is therefore associated with spatial derivatives of the scalar field, and can be written as

$$V' = \frac{1}{2}\sum_i (\partial_i \phi)^2 = \frac{1}{2}(\nabla\phi)^2\,, \tag{10.4}$$

where ∂_i are derivatives with respect to spatial coordinates. We have written this term as a new contribution V' to the potential energy, rather than as a contribution to the kinetic energy. There are several reasons for this. First, it is needed for Lorentz invariance. The two options lead to contributions of opposite signs in the Lagrangian, and only one sign is consistent with Lorentz invariance. Second, kinetic energy is always associated with time derivatives. Third, the calculation of the total energy vindicates the correctness of the choice. Indeed, with this additional term the Lagrangian density becomes

$$\mathcal{L} = T - V' - V = \frac{1}{2}\,\partial_0\phi\,\partial_0\phi - \frac{1}{2}\,\partial_i\phi\,\partial_i\phi - \frac{1}{2}\,m^2\phi^2\,, \tag{10.5}$$

where the repeated spatial index i denotes summation. The relative sign between the first two terms on the right-hand side allows us to rewrite them as a single term which uses the Minkowski metric $\eta^{\mu\nu}$:

$$\mathcal{L} = -\frac{1}{2}\,\eta^{\mu\nu}\partial_\mu\phi\,\partial_\nu\phi - \frac{1}{2}\,m^2\phi^2\,. \tag{10.6}$$

Since all the indices are matched, the Lagrangian density is Lorentz invariant. The associated action is

$$S = \int d^D x \left(-\frac{1}{2}\,\eta^{\mu\nu}\,\partial_\mu\phi\partial_\nu\phi - \frac{1}{2}m^2\phi^2\right), \tag{10.7}$$

where $d^D x = dx^0 dx^1 \ldots dx^d$, and $D = d + 1$, is the number of spacetime dimensions. This is the action for a *free* scalar field with mass m. A field is said to be free when its equations of motion are linear. If each term in the action is quadratic in the field, as is the case in (10.7), the equations of motion will be linear in the field. A field that is not free is said to be interacting, in which case the action contains terms of order three or higher in the field.

To find the energy density in this field we calculate the Hamiltonian density $\mathcal{H}$. The momentum Π conjugate to the field is given by

$$\Pi \equiv \frac{\partial\mathcal{L}}{\partial(\partial_0\phi)} = \partial_0\phi\,, \tag{10.8}$$

where we used (10.5) to evaluate the derivative. The Hamiltonian density is then constructed as

$$\mathcal{H} = \Pi\partial_0\phi - \mathcal{L}. \tag{10.9}$$

Quick calculation 10.1 Show that the Hamiltonian density takes the form

$$\mathcal{H} = \frac{1}{2}\,\Pi^2 + \frac{1}{2}\,(\nabla\phi)^2 + \frac{1}{2}\,m^2\phi^2. \tag{10.10}$$

The three terms in $\mathcal{H}$ are identified as T, V', and V, respectively. This is what we expected physically for the energy density. The total energy E is given by the Hamiltonian H, which in turn, is the spatial integral of the Hamiltonian density $\mathcal{H}$:

$$E = H = \int d^d x \Big(\frac{1}{2}\,\partial_0\phi\,\partial_0\phi + \frac{1}{2}\,(\nabla\phi)^2 + \frac{1}{2}\,m^2\phi^2 \Big)\,. \tag{10.11}$$

To find the equations of motion from the action (10.7), we consider a variation $\delta\phi$ of the field and set the variation of the action equal to zero. After discarding a total derivative we find

$$\delta S = \int d^D x \Big(-\eta^{\mu\nu}\partial_\mu(\delta\phi)\partial_\nu\phi - m^2\phi\delta\phi \Big) = \int d^D x\; \delta\phi \Big(\eta^{\mu\nu}\partial_\mu\partial_\nu\phi - m^2\phi \Big) = 0\,. \tag{10.12}$$

The equation of motion for ϕ is therefore

$$\eta^{\mu\nu}\partial_\mu\partial_\nu\phi - m^2\phi = 0\,. \tag{10.13}$$

If we define $\partial^2 \equiv \eta^{\mu\nu}\partial_\mu\partial_\nu$, then we have

$$(\partial^2 - m^2)\,\phi = 0\,. \tag{10.14}$$

Separating out time and space derivatives, this equation is recognized as the Klein–Gordon equation:

$$-\frac{\partial^2\phi}{\partial t^2} + \nabla^2\phi - m^2\phi = 0\,. \tag{10.15}$$

We will now study some classical solutions of this equation.

10.3 Classical plane-wave solutions

We can find plane-wave solutions to the classical scalar field equation (10.15). Consider, for example, the expression

$$\phi(t,\vec{x}) = a\, e^{-iEt + i\vec{p}\cdot\vec{x}}\,, \tag{10.16}$$

where a and E are constants and $\vec{p}$ is an arbitrary vector. The field equation (10.15) fixes the possible values of E in terms of $\vec{p}$ and m:

$$E^2 - \vec{p}^{\,2} - m^2 = 0 \;\longrightarrow\; E = \pm E_p\,, \quad E_p \equiv \sqrt{\vec{p}^{\,2} + m^2}\,. \tag{10.17}$$

The square root is chosen to be positive, so $E_p > 0$. There is a small problem with the solution in (10.16). While ϕ is a real field, the solution (10.16) is not real. To make it real, we just add to it its complex conjugate:

$$\phi(t, \vec{x}) = a\, e^{-iE_p t + i\vec{p}\cdot\vec{x}} + a^*\, e^{iE_p t - i\vec{p}\cdot\vec{x}}\,. \tag{10.18}$$

This solution depends on the complex number a. A general solution to the equation of motion (10.15) is obtained by superimposing solutions, such as those above, for all values of $\vec{p}$. Since $\vec{p}$ can be varied continuously, the general superposition is actually an integral. The classical field does not have a simple quantum mechanical interpretation. If the two terms above were to be thought of as wavefunctions, the first would represent the wavefunction of a particle with momentum $\vec{p}$ and positive energy E_p. The second would represent the wavefunction of a particle with momentum $(-\vec{p})$ and *negative* energy $(-E_p)$. This is not acceptable. To do quantum mechanics with a classical field one must quantize the field. The result is particle states with positive energy, as we will discuss briefly in the following section.

An analysis of the classical field equation (in a practical way that applies elsewhere) uses the Fourier transformation of the scalar field $\phi(x)$:

$$\phi(x) = \int \frac{d^D p}{(2\pi)^D}\; e^{ip\cdot x}\, \phi(p)\,. \tag{10.19}$$

Here $\phi(p)$ is the Fourier transform of $\phi(x)$. We will always show the argument of ϕ so no confusion should arise between the spacetime field and the momentum-space field. Note that we are performing the Fourier transform over all spacetime coordinates, time included: $p\cdot x = -p^0 x^0 + \vec{p}\cdot\vec{x}$. The reality of $\phi(x)$ means that $\phi(x) = (\phi(x))^*$. Using equation (10.19), this condition yields

$$\int \frac{d^D p}{(2\pi)^D}\; e^{ip\cdot x}\, \phi(p) = \int \frac{d^D p}{(2\pi)^D}\; e^{-ip\cdot x}\, (\phi(p))^*\,. \tag{10.20}$$

We let $p \to -p$ on the left-hand side of this equation. This change of integration variable does not affect the integration $\int d^D p$, and results in

$$\int \frac{d^D p}{(2\pi)^D}\; e^{-ip\cdot x} \Big(\phi(-p) - (\phi(p))^*\Big) = 0\,, \tag{10.21}$$

where we collected all terms on the left-hand side. This left-hand side is a function of x that must vanish identically. It is also the Fourier transform of the momentum-space function in parentheses. This function must therefore vanish:

$$(\phi(p))^* = \phi(-p)\,. \tag{10.22}$$

This is the reality condition in momentum space.

Substituting (10.19) into (10.14) and letting ∂^2 act on $e^{ip\cdot x}$ we find

$$\int \frac{d^D p}{(2\pi)^D}\; (-p^2 - m^2)\, \phi(p)\, e^{ipx} = 0\,. \tag{10.23}$$

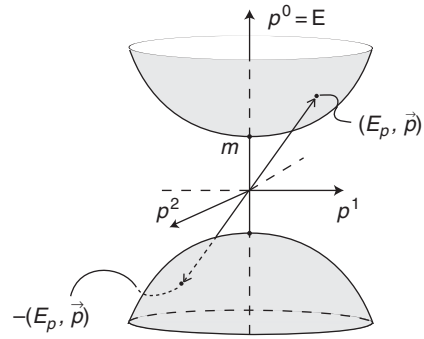

Fig. 10.1 **The mass-shell hyperboloid $E^2 - p^2 = m^2$ (only two spatial directions shown). The equation of motion sets the mass m scalar field $\phi(p)$ equal to zero away from the hyperboloid. On the hyperboloid the field is arbitrary up to a reality condition that relates the field values at antipodal points.**

Since this must hold for all values of x, this equation requires

$$(p^2 + m^2)\,\phi(p) = 0 \quad \text{for all} \;\; p\,. \tag{10.24}$$

This is a simple equation: $(p^2 + m^2)$ is just a number multiplying $\phi(p)$ and the product must vanish. Solving this equation means specifying the values of $\phi(p)$ for all values of p. Since either factor may vanish we must consider two cases.

(i) $p^2 + m^2 \neq 0$. In this case the scalar field vanishes: $\phi(p) = 0$.
(ii) $p^2 + m^2 = 0$. In this case the scalar field $\phi(p)$ is arbitrary.

In momentum space the hypersurface $p^2 + m^2 = 0$ is called the *mass-shell*. With $p^\mu = (E, \vec{p}\,)$, the mass-shell is the locus of points in momentum space where $E^2 = \vec{p}^{\,2} + m^2$, the hyperboloid sketched in Figure 10.1. The mass-shell is therefore the set of points $(\pm E_p, \vec{p})$, for all values of $\vec{p}$. We have learned that $\phi(p)$ vanishes off the mass-shell and is arbitrary (up to the reality condition) on the mass-shell.

We now introduce the idea of *classical degrees of freedom*. For a point p^μ on the mass-shell, the solution is determined by specifying the complex number $\phi(p)$. This number determines as well the solution at the point $(-p^\mu)$, also belonging to the mass-shell: $\phi(-p) = (\phi(p))^*$. So, a complex number fixes the values of the field for *two* points on the mass-shell. We need, on average, one real number for each point on the mass-shell. We will say that a field satisfying equation (10.24) represents *one degree of freedom per point on the mass-shell*.

We conclude this section by writing the scalar field equation of motion in light-cone coordinates. Let $\vec{x}_T$ denote a vector whose components are the transverse coordinates x^I:

$$\vec{x}_T = (x^2, x^3, \ldots, x^d)\,. \tag{10.25}$$

In this notation, the collection of spacetime coordinates becomes $(x^+, x^-, \vec{x}_T)$. Equation (10.14) expanded in light-cone coordinates is

$$\left(-2\frac{\partial}{\partial x^+}\frac{\partial}{\partial x^-} + \frac{\partial}{\partial x^I}\frac{\partial}{\partial x^I} - m^2\right)\phi(x^+, x^-, \vec{x}_T) = 0\,. \tag{10.26}$$

To simplify this equation, we Fourier transform the *spatial* dependence of the field, changing x^- into p^+ and x^I into p^I. Letting $\vec{p}_T$ denote the vector whose components are the transverse momenta p^I

$$\vec{p}_T = (p^2, p^3, \ldots, p^d)\,, \tag{10.27}$$

the Fourier transform is written as

$$\phi(x^+, x^-, \vec{x}_T) = \int \frac{dp^+}{2\pi} \int \frac{d^{D-2}\vec{p}_T}{(2\pi)^{D-2}}\, e^{-ix^- p^+ + i\,\vec{x}_T\cdot\vec{p}_T}\, \phi(x^+, p^+, \vec{p}_T)\,. \tag{10.28}$$

We now substitute this form of the scalar field into (10.26) to get

$$\left(-2\frac{\partial}{\partial x^+}(-ip^+) - p^I p^I - m^2\right)\phi(x^+, p^+, \vec{p}_T) = 0\,, \tag{10.29}$$

and dividing by $2p^+$ we find

$$\left(i\,\frac{\partial}{\partial x^+} - \frac{1}{2p^+}(p^I p^I + m^2)\right)\phi(x^+, p^+, \vec{p}_T) = 0\,. \tag{10.30}$$

This is the equation we were after. As opposed to the original Lorentz covariant equation of motion, which has second-order derivatives with respect to time, the light-cone equation is a first-order differential equation in light-cone time. Equation (10.30) has the formal structure of a Schrödinger equation, which is also first order in time. This fact will prove useful when we study how the quantum point particle is related to the scalar field.

Another version of equation (10.30) will be needed in our later work. Using a new time parameter τ related to x^+ as $x^+ = p^+\tau/m^2$ we obtain

$$\left(i\,\frac{\partial}{\partial \tau} - \frac{1}{2m^2}(p^I p^I + m^2)\right)\phi(\tau, p^+, \vec{p}_T) = 0\,. \tag{10.31}$$

Quick calculation 10.2 Consider the mass-shell condition $p^2 + m^2 = 0$ in light-cone coordinates. Show that

$$p^- = \frac{1}{2p^+}(p^I p^I + m^2)\,. \tag{10.32}$$

10.4 Quantum scalar fields and particle states

Quantum field theory is a natural language to describe the quantum behavior of elementary particles and their interactions. Quantum field theory is quantum mechanics applied to classical fields. In quantum mechanics, classical dynamical variables turn into operators. The position and momentum of a classical particle, for example, turn into position and momentum operators. If our dynamical variables are classical fields, the quantum operators will

be *field operators*. Thus, in quantum field theory, the fields are operators. The state space in a quantum field theory is typically described using a set of *particle states*. Quantum field theory also has energy and momentum operators. When they act on a particle state, these operators give the energy and momentum of the particles described by the state.

In this section, we examine briefly how the features just described arise concretely. Inspired by the plane-wave solution (10.18), which describes a superposition of complex waves with momenta $\vec{p}$ and $-\vec{p}$ (with $\vec{p} \neq 0$), we consider a classical field configuration $\phi_p(t, \vec{x})$ of the form

$$\phi_p(t, \vec{x}) = \frac{1}{\sqrt{V}} \frac{1}{\sqrt{2E_p}} \Big(a(t)\, e^{i\vec{p}\cdot\vec{x}} + a^*(t)\, e^{-i\vec{p}\cdot\vec{x}} \Big) . \tag{10.33}$$

There have been two changes. First, the time dependence has been made more general by introducing the function $a(t)$ and its complex conjugate $a^*(t)$. The function $a(t)$ is the dynamical variable that determines the field configuration. Second, we have placed a normalization factor $\sqrt{V}$, where V is assumed to be the volume of space. The normalization factor also includes a square root of the energy E_p defined in equation (10.17).

We can imagine space as a box with sides $L_1, L_2, \ldots, L_d$, in which case $V = L_1 L_2 \ldots L_d$. When we put a field on a box we usually require it to be periodic. The field ϕ_p is periodic if each component p_i of $\vec{p}$ satisfies

$$p_i L_i = 2\pi n_i , \quad i = 1, 2, \ldots, d . \tag{10.34}$$

Here the n_i are integers. Each component of the momentum is quantized.

We now try to do quantum mechanics with the field configuration (10.33). To this end, we evaluate the scalar field action (10.7) for $\phi = \phi_p(t, \vec{x})$:

$$S = \int dt \int d^d x \Big(\frac{1}{2} (\partial_0 \phi_p)^2 - \frac{1}{2} (\nabla \phi_p)^2 - \frac{1}{2} m^2 \phi_p^2 \Big) . \tag{10.35}$$

The evaluation involves squaring the field, squaring its time derivative, and squaring its gradient. In squaring any of these, we obtain from the cross multiplication two types of terms: those with spatial dependence $\exp(\pm 2i\, \vec{p} \cdot \vec{x})$ and those without spatial dependence. We claim that the spatial integral $\int d^d x$ of the terms with spatial dependence is zero, so these terms cannot contribute. Indeed, the quantization conditions in (10.34) imply that

$$\int_0^{L_1} dx^1 \ldots \int_0^{L_d} dx^d \, \exp(\pm 2i\, \vec{p} \cdot \vec{x}) = 0 . \tag{10.36}$$

For the terms without spatial dependence, the spatial integral gives a factor of the volume V, which cancels with the product of $\sqrt{V}$ factors we introduced in (10.33). The result is

$$S = \int dt \, \Big(\frac{1}{2E_p} \dot{a}^*(t) \dot{a}(t) - \frac{1}{2} E_p \, a^*(t) a(t) \Big) . \tag{10.37}$$

Quick calculation 10.3 Verify that equation (10.37) is correct.

Similarly, we use equation (10.11) to evaluate the field energy H:

$$H = \frac{1}{2E_p} \dot{a}^*(t) \dot{a}(t) + \frac{1}{2} E_p \, a^*(t) a(t) . \tag{10.38}$$

Quick calculation 10.4 Verify that equation (10.38) is correct.

The action (10.37) describes the dynamics of two simple harmonic oscillators. Since $a(t)$ is a complex dynamical variable we can write

$$a(t) = q_1(t) + i\, q_2(t), \tag{10.39}$$

where $q_1(t)$ and $q_2(t)$ are real coordinates. In terms of $q_1(t)$ and $q_2(t)$, the action becomes

$$S = \int dt L = \sum_{i=1}^{2} \int dt \Big(\frac{1}{2E_p}\, \dot{q}_i^2(t) - \frac{1}{2}\, E_p\, q_i^2(t) \Big) . \tag{10.40}$$

We see here that $q_1(t)$ and $q_2(t)$ are indeed the coordinates of identical simple harmonic oscillators. Their canonical momenta are

$$p_i(t) = \frac{\partial L}{\partial \dot{q}_i} = \frac{\dot{q}_i(t)}{E_p} \;\longrightarrow\; p_1(t) + i p_2(t) = \frac{1}{E_p}\, \dot{a}(t) . \tag{10.41}$$

The equations of motion that follow from the variation of the action (10.40) are

$$\ddot{q}_i(t) = -E_p^2\, q_i(t) , \quad i = 1, 2. \tag{10.42}$$

It is actually convenient to work with the complex combination $a(t)$. Using (10.39) the equations of motion become the single complex equation

$$\ddot{a}(t) = -E_p^2\, a(t) . \tag{10.43}$$

Equation (10.43) is readily solved in terms of exponentials. Since it is a second-order equation, there are two solutions:

$$a(t) = a_p\, e^{-iE_p t} + a_{-p}^*\, e^{iE_p t} . \tag{10.44}$$

There is no reality condition since $a(t)$ is complex. In writing the above solution we introduced two independent complex constants a_p and a_{-p}^*. Substituting this solution into (10.38), we find

$$H = E_p \left(a_p^* a_p + a_{-p}^* a_{-p} \right) . \tag{10.45}$$

Note that the time dependence disappeared; the energy is conserved.

In classical scalar field theory there is an integral expression that gives the spacetime momentum $\vec{P}$ carried by the field. In the present case, this field momentum is conserved, and a formula can be obtained from the analysis of Problem 8.10. We will not discuss this here, but the answer is

$$\vec{P} = -\int d^d x (\partial_0 \phi) \nabla \phi . \tag{10.46}$$

In Problem 10.1 you are asked to evaluate $\vec{P}$ for the field configuration (10.33), when $a(t)$ is given by (10.44). The result that you will find is

$$\vec{P} = \vec{p} \left(a_p^* a_p - a_{-p}^* a_{-p} \right) . \tag{10.47}$$

The system at hand consists of two harmonic oscillators. The expression for H suggests that in the quantum theory a_p and a_{-p} become annihilation operators, while a_p^* and a_{-p}^* become the creation operators $a_p^\dagger$ and $a_{-p}^\dagger$, respectively. We will verify the correctness of this assumption shortly. These oscillators are required to satisfy the standard commutation relations:

$$[\, a_p \,,\, a_p^\dagger \,] = 1 \,, \quad [\, a_{-p} \,,\, a_{-p}^\dagger \,] = 1 \,. \tag{10.48}$$

Any commutator which involves an operator with subscript p and an operator with subscript $(-p)$ is declared to vanish. The variable $a(t)$ in (10.44) now becomes an operator. We record its form, together with that for its Hermitian conjugate $a^\dagger(t)$, as well as those for their time derivatives:

$$\begin{aligned} a(t) &= a_p\, e^{-iE_p t} + a_{-p}^\dagger\, e^{iE_p t} \,, \\ a^\dagger(t) &= a_p^\dagger\, e^{iE_p t} + a_{-p}\, e^{-iE_p t} \,, \\ \dot{a}(t) &= -iE_p \left(a_p\, e^{-iE_p t} - a_{-p}^\dagger\, e^{iE_p t} \right) , \\ \dot{a}^\dagger(t) &= iE_p \left(a_p^\dagger\, e^{iE_p t} - a_{-p}\, e^{-iE_p t} \right) . \end{aligned} \tag{10.49}$$

It takes a short computation to verify that the only nonvanishing commutators among $a(t)$, $a^\dagger(t)$, $\dot{a}(t)$, and $\dot{a}^\dagger(t)$ are

$$\Big[\, a(t) \,,\, \dot{a}^\dagger(t) \,\Big] = \Big[\, a^\dagger(t) \,,\, \dot{a}(t) \,\Big] = 2iE_p \,. \tag{10.50}$$

We can now show that these commutation relations imply the commutation relations that we would naturally impose between the coordinates and their corresponding conjugate momenta:

$$[\, q_1(t),\, p_1(t) \,] = [\, q_2(t),\, p_2(t) \,] = i \,. \tag{10.51}$$

To do this, we can use equations (10.39) and (10.41) to solve for the coordinates and the momenta:

$$q_1(t) = \frac{1}{2} \left(a(t) + a^\dagger(t) \right) , \qquad p_1(t) = \frac{1}{2E_p} \left(\dot{a}(t) + \dot{a}^\dagger(t) \right) , \tag{10.52}$$

$$q_2(t) = \frac{1}{2i} \left(a(t) - a^\dagger(t) \right) , \qquad p_2(t) = \frac{1}{2iE_p} \left(\dot{a}(t) - \dot{a}^\dagger(t) \right) . \tag{10.53}$$

We can now check that, for example,

$$[\, q_1(t),\, p_1(t) \,] = \frac{1}{4E_p} \Big[\, a(t) + a^\dagger(t) \,,\, \dot{a}(t) + \dot{a}^\dagger(t) \,\Big] = \frac{1}{4E_p} (2iE_p + 2iE_p) = i \,, \tag{10.54}$$

as expected. The other commutators can be checked similarly, thus confirming that the postulated commutation relations (10.48) are correct.

Quick calculation 10.5 Check that $[\, q_2(t),\, p_2(t) \,] = i$ and that $[\, q_1(t),\, p_2(t) \,] = 0$.

At the quantum level, the Hamiltonian (10.45) becomes the operator

$$H = E_p \Big(a_p^\dagger a_p + a_{-p}^\dagger a_{-p} \Big) \,, \tag{10.55}$$

describing a pair of simple harmonic oscillators, each of which has frequency E_p. The momentum (10.46) becomes the operator

$$\vec{P} = \vec{p}\left(a_p^\dagger a_p - a_{-p}^\dagger a_{-p}\right). \tag{10.56}$$

Note that the oscillators with subscripts $(-p)$ contribute with a negative sign to the momentum. This equation will help our interpretation below.

The first two equations in (10.49) can be used back in (10.33) to obtain the operator version of the field configuration:

$$\begin{aligned}\phi_p(t,\vec{x}) = &\frac{1}{\sqrt{V}}\frac{1}{\sqrt{2E_p}}\left(a_p\, e^{-iE_pt+i\vec{p}\cdot\vec{x}} + a_p^\dagger\, e^{iE_pt-i\vec{p}\cdot\vec{x}}\right)\\ &+\frac{1}{\sqrt{V}}\frac{1}{\sqrt{2E_p}}\left(a_{-p}\, e^{-iE_pt-i\vec{p}\cdot\vec{x}} + a_{-p}^\dagger\, e^{iE_pt+i\vec{p}\cdot\vec{x}}\right).\end{aligned} \tag{10.57}$$

We see that ϕ_p is in fact a spacetime dependent operator, or a field operator. The second line in (10.57) is obtained from the first line by the replacement $\vec{p} \to -\vec{p}$, which does not affect E_p. In full generality, the quantum field $\phi(x)$ includes contributions from all values of the spatial momentum $\vec{p}$:

$$\phi(t,\vec{x}) = \frac{1}{\sqrt{V}}\sum_{\vec{p}}\frac{1}{\sqrt{2E_p}}\left(a_p\, e^{-iE_pt+i\vec{p}\cdot\vec{x}} + a_p^\dagger\, e^{iE_pt-i\vec{p}\cdot\vec{x}}\right). \tag{10.58}$$

The commutation relations for the oscillators are the natural generalization of those in (10.48):

$$[a_p\,, a_k^\dagger] = \delta_{p,k}\,, \quad [a_p\,, a_k] = [a_p^\dagger\,, a_k^\dagger] = 0. \tag{10.59}$$

All subscripts here are spatial vectors, written without the arrows to avoid cluttering the equations. The Kronecker delta $\delta_{p,k}$ is zero unless $\vec{p} = \vec{k}$, in which case it equals one. Once we consider contributions from all values of the momenta, the previous expression for the Hamiltonian in (10.55) and that for the momentum operator in (10.56) must be changed. One can show that

$$H = \sum_{\vec{p}} E_p\, a_p^\dagger a_p\,, \tag{10.60}$$

$$\vec{P} = \sum_{\vec{p}} \vec{p}\, a_p^\dagger a_p\,. \tag{10.61}$$

We will not derive these expressions, but they should seem quite plausible.

The state space of this quantum system is built in the same way as the state space of the simple harmonic oscillator. We assume the existence of a vacuum state $|\Omega\rangle$, which acts just

as the simple harmonic oscillator ground state $|0\rangle$ in that it is annihilated by the annihilation operators a_p: $a_p|\Omega\rangle = 0$ for all $\vec{p}$. It follows that $H|\Omega\rangle = 0$, which makes the vacuum a zero-energy state. This vacuum state is interpreted as a state in which there are no particles. On the other hand, the state

$$a_p^\dagger\,|\Omega\rangle \tag{10.62}$$

is interpreted as a state with precisely one particle. We claim that it is a particle with momentum $\vec{p}$. To verify this, we act on the state with the momentum operator (10.61) and use (10.59) to find

$$\vec{P}a_p^\dagger\,|\Omega\rangle = \sum_{\vec{k}} \vec{k}a_k^\dagger\,[a_k\,,\,a_p^\dagger]\,|\Omega\rangle = \vec{p}\,a_p^\dagger\,|\Omega\rangle\,. \tag{10.63}$$

The energy of the state is similarly computed by acting on the state with the Hamiltonian H:

$$Ha_p^\dagger\,|\Omega\rangle = \sum_{\vec{k}} E_k\,a_k^\dagger\,[a_k\,,\,a_p^\dagger]\,|\Omega\rangle = E_p\,a_p^\dagger\,|\Omega\rangle\,. \tag{10.64}$$

The state $a_p^\dagger|\Omega\rangle$ has positive energy. While the quantum field operator has both positive and negative energy components, the states that represent the particles have positive energy. The states $a_p^\dagger\,|\Omega\rangle$ are the *one-particle* states.

The state space contains multiparticle states as well. These are states built by acting on the vacuum with a collection of creation operators:

$$a_{p_1}^\dagger\,a_{p_2}^\dagger\ldots a_{p_k}^\dagger|\Omega\rangle\,. \tag{10.65}$$

This state, with k creation operators acting on the vacuum, represents a state with k particles. The particles have momenta $\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_k$ and energies $E_{p_1}, E_{p_2}, \ldots, E_{p_k}$. The various momenta $\vec{p}_i$ need not be all different.

Quick calculation 10.6 Show that the eigenvalues of $\vec{P}$ and H acting on (10.65) are $\sum_{n=1}^k \vec{p}_n$ and $\sum_{n=1}^k E_{p_n}$, respectively.

Quick calculation 10.7 Convince yourself that $N = \sum_{\vec{p}} a_p^\dagger a_p$ is a number operator: acting on a state it gives us the number of particles contained in the state.

Our analysis of classical solutions in the previous section led to the conclusion that there is one degree of freedom per point on the mass-shell. At the quantum level, we focus on the one-particle states. Consequently, we restrict ourselves to the physical part of the mass-shell, the part where the energy is positive ($p^0 = E > 0$). We have a single one-particle state for each point on the physical mass-shell. This state is labeled by its spatial momentum $\vec{p}$.

To describe the particle states in light-cone coordinates, the changes are minimal. The physical part of the mass-shell is parameterized by the transverse momenta $\vec{p}_T$ and the light-cone momenta p^+ for which $p^+ > 0$. The value of the light-cone energy p^- is then

fixed. Thus, instead of labeling the oscillators with $\vec{p}$, we simply label them with p^+ and $\vec{p}_T$. The one-particle states are written as

$$\text{One-particle states of a scalar field:} \quad a^\dagger_{p^+,\,p_T}\,|\Omega\rangle. \tag{10.66}$$

The momentum operator given in (10.61) has a natural light-cone version. The various components of the operator take the form

$$\begin{aligned}
\hat{p}^+ &= \sum_{p^+,p_T} p^+\, a^\dagger_{p^+,\,p_T} a_{p^+,\,p_T}\,,\\
\hat{p}^I &= \sum_{p^+,p_T} p^I\, a^\dagger_{p^+,\,p_T} a_{p^+,\,p_T}\,,\\
\hat{p}^- &= \sum_{p^+,p_T} \frac{1}{2p^+}\Big(p^I p^I + m^2\Big)\, a^\dagger_{p^+,\,p_T} a_{p^+,\,p_T}\,.
\end{aligned} \tag{10.67}$$

In the last equation, the factor which multiplies the oscillators is the value of p^- determined from the mass-shell condition (10.32). This last equation is analogous to (10.60), where E_p is the energy determined from the mass-shell condition.

10.5 Maxwell fields and photon states

We now turn to an analysis of Maxwell fields and their corresponding quantum states. As opposed to the case of the scalar field, where there is no gauge invariance, electromagnetic fields have a gauge invariance that will make our analysis more subtle and interesting. In order to study the field equations in a convenient way, we will impose the gauge condition that defines the light-cone gauge. We will then be able to describe the quantum states of the Maxwell field.

The field equations for electromagnetism are written in terms of the electromagnetic vector potential $A_\mu(x)$. As we reviewed in Section 3.3, the field strength $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is invariant under the gauge transformation

$$\delta A_\mu = \partial_\mu \epsilon\,, \tag{10.68}$$

where ϵ is the gauge parameter. The field equations take the form

$$\partial_\nu F^{\mu\nu} = 0 \;\longrightarrow\; \partial_\nu(\partial^\mu A^\nu - \partial^\nu A^\mu) = 0\,, \tag{10.69}$$

and can be written as

$$\partial^2 A^\mu - \partial^\mu(\partial\cdot A) = 0\,. \tag{10.70}$$

Compare this equation with equation (10.14) for the scalar field. There is no indication of a mass term for the Maxwell field – such a term would be recognized as one without spacetime derivatives. We will confirm below that the Maxwell field is, indeed, massless.

We Fourier transform all the components of the vector potential in order to determine the equation of motion in momentum space:

$$A^\mu(x) = \int \frac{d^D p}{(2\pi)^D}\, e^{ipx} A^\mu(p)\,, \tag{10.71}$$

where reality of $A^\mu(x)$ implies $A^\mu(-p) = (A^\mu(p))^*$. Substituting (10.71) into (10.70) we obtain the equation:

$$p^2 A^\mu - p^\mu (p \cdot A) = 0\,. \tag{10.72}$$

The gauge transformation (10.68) can also be Fourier transformed. In momentum space, the gauge transformation relates $\delta A_\mu(p)$ to the Fourier transform $\epsilon(p)$ of the gauge parameter:

$$\delta A_\mu(p) = i p_\mu\, \epsilon(p)\,. \tag{10.73}$$

Since the gauge parameter $\epsilon(x)$ is real, we have $\epsilon(-p) = \epsilon^*(p)$. The gauge transformation (10.73) is consistent with the reality of $\delta A_\mu(x)$. Indeed,

$$(\delta A_\mu(p))^* = -i p_\mu (\epsilon(p))^* = i(-p_\mu)\epsilon(-p) = \delta A_\mu(-p)\,. \tag{10.74}$$

Note the role of the factor of i in getting the signs to work out.

Being done with preliminaries, we can analyze (10.72) subject to the gauge transformations (10.73). At this point, it is more convenient to work with the light-cone components of the gauge field:

$$A^+(p)\,, \quad A^-(p)\,, \quad A^I(p)\,. \tag{10.75}$$

The gauge transformations (10.73) then read

$$\delta A^+ = i p^+ \epsilon\,, \quad \delta A^- = i p^- \epsilon\,, \quad \delta A^I = i p^I \epsilon\,. \tag{10.76}$$

We now impose a gauge condition. As we have emphasized before, when working with the light-cone formalism we always assume $p^+ \neq 0$. The above gauge transformations now make it clear that we can set A^+ to zero by choosing ϵ correctly. Indeed, if we apply a gauge transformation

$$A^+ \to A'^+ = A^+ + i p^+ \epsilon\,, \tag{10.77}$$

then the $+$ component of the new gauge field A' vanishes if we choose $\epsilon = i A^+/p^+$. In other words, we can always make the $+$ component of the Maxwell field zero by applying a gauge transformation. This will be our defining condition for the light-cone gauge in Maxwell theory:

$$\text{light-cone gauge condition}: \quad A^+(p) = 0. \tag{10.78}$$

Setting A^+ to zero determines the gauge parameter ϵ, and no additional gauge transformations are possible: if $A^+ = 0$, any further gauge transformation will make it nonzero.

There is one minor exception: any parameter of the form $\epsilon(p) = \epsilon(p^-, p^I)\delta(p^+)$ will keep $A^+ = 0$ because $p^+\epsilon(p) = 0$. The similarities with light-cone gauge string theory are noteworthy. In light-cone gauge open string theory all world-sheet reparameterization invariances are fixed. Moreover, while not equal to zero, X^+ is very simple: the corresponding zero mode and oscillators vanish.

The gauge condition (10.78) simplifies the equation of motion (10.72) considerably. Taking $\mu = +$ we find

$$p^+(p \cdot A) = 0 \longrightarrow p \cdot A = 0\,. \tag{10.79}$$

This equation can be expanded out using light-cone indices:

$$-p^+A^- - p^-A^+ + p^I A^I = 0\,. \tag{10.80}$$

Since $A^+ = 0$, this equation determines A^- in terms of the transverse A^I:

$$A^- = \frac{1}{p^+}(p^I A^I)\,. \tag{10.81}$$

This is reminiscent of our light-cone analysis of the string, where X^- was solved for in terms of the transverse coordinates (and a zero mode). Using (10.79) back in (10.72), all that remains from the field equation is

$$p^2 A^\mu(p) = 0\,. \tag{10.82}$$

For $\mu = +$ this equation is trivially satisfied, since $A^+ = 0$. For $\mu = I$ we get a set of nontrivial conditions:

$$p^2 A^I(p) = 0\,. \tag{10.83}$$

For $\mu = -$, we get $p^2A^-(p) = 0$. This is automatically satisfied on account of (10.81) and (10.83).

For each value of I, equation (10.83) takes the form of the equation of motion for a massless scalar. Thus, $A^I(p) = 0$ when $p^2 \neq 0$. This makes $A^- = 0$, and since A^+ is zero, the full gauge field vanishes. For $p^2 = 0$, the $A^I(p)$ are unconstrained, and the $A^-(p)$ are determined as a function of the A^I (see (10.81)). The degrees of freedom of the Maxwell field are thus carried by the $(D-2)$ transverse fields $A^I(p)$, for $p^2 = 0$. We say that we have $(D-2)$ degrees of freedom per point on the mass-shell.

It is actually possible to show that there are no degrees of freedom for $p^2 \neq 0$, without having to make a choice of gauge. Although not every field is zero, every field is *gauge equivalent* to the zero field when $p^2 \neq 0$. If a field differs from the zero field by only a gauge transformation, we say that the field is *pure gauge*. Recall that fields A_μ and A'_μ are gauge equivalent if $A_\mu = A'_\mu + \partial_\mu\chi$, for some scalar function χ. Taking $A'_\mu = 0$, we learn that $A_\mu = \partial_\mu\chi$ is gauge equivalent to the zero field, and is therefore pure gauge. The term pure gauge is suitable: A_μ takes the form of a gauge transformation. In momentum space, a pure gauge is a field that can be written as

$$\text{pure gauge:} \quad A_\mu(p) = ip_\mu\chi(p)\,, \tag{10.84}$$

for some choice of χ. Rewrite now the equation of motion (10.72) as

$$p^2 A_\mu = p_\mu (p \cdot A) \,. \tag{10.85}$$

Since $p^2 \neq 0$, we can write

$$A_\mu = i p_\mu \Big(\frac{-ip \cdot A}{p^2} \Big) \,. \tag{10.86}$$

Comparing with (10.84), we see that A_μ is pure gauge. This means that there are no degrees of freedom in the Maxwell field when $p^2 \neq 0$. For all intents and purposes, there is no field.

Let us now discuss briefly photon states. Each of the independent classical fields A^I can be expanded as we did for the scalar field in (10.58). To do so, we introduce – as you can infer by analogy – oscillators a_p^I and $a_p^{I\dagger}$, where the subscripts p represent the values of p^+ and $\vec{p}_T$. We thus get $(D-2)$ species of oscillators. Introducing a vacuum $|\Omega\rangle$, the one-photon states would be written as

$$a^{I\dagger}_{p^+,\, p_T} |\Omega\rangle \,. \tag{10.87}$$

Here the label I is a polarization label. The photon state (10.87) is said to be polarized in the Ith direction. Since we have $(D-2)$ possible polarizations, *we have* $(D-2)$ *linearly independent one-photon states for each point on the physical sector of the mass-shell*. A general one-photon state with momentum $(p^+, \vec{p}_T)$ is a linear superposition of the above states:

$$\text{one-photon states:} \quad \sum_{I=2}^{D-1} \xi_I \, a^{I\dagger}_{p^+,\, p_T} |\Omega\rangle . \tag{10.88}$$

Here the transverse vector ξ_I is called the polarization vector.

In four-dimensional spacetime, Maxwell theory gives rise to $D-2=2$ single-photon states for any fixed spatial momentum. This is indeed familiar to you, at least classically. An electromagnetic plane wave which propagates in a fixed direction and has some fixed wavelength (i.e., fixed momentum), can be written as a superposition of two plane waves that represent independent polarization states.

10.6 Gravitational fields and graviton states

Gravitation emerges in string theory in the language of Einstein's theory of general relativity. We discussed this language briefly in Section 3.6. The dynamical field variable is the spacetime metric $g_{\mu\nu}(x)$, which in the approximation of weak gravitational fields can be taken to be of the form $g_{\mu\nu}(x) = \eta_{\mu\nu} + h_{\mu\nu}(x)$. Both $g_{\mu\nu}$ and $h_{\mu\nu}$ are symmetric under the exchange of their indices. The field equations for $g_{\mu\nu}$ – Einstein's equations – can be used to derive a linearized equation of motion for the fluctuations $h_{\mu\nu}$. This equation was given

in (3.82). Defining $h_{\mu\nu}(p)$ to be the Fourier transform of $h_{\mu\nu}(x)$, the momentum-space version of this equation is

$$S^{\mu\nu}(p) \equiv p^2 h^{\mu\nu} - p_\alpha (p^\mu h^{\nu\alpha} + p^\nu h^{\mu\alpha}) + p^\mu p^\nu h = 0\,. \tag{10.89}$$

If we were considering Einstein's equations in the presence of sources, the right-hand side of this equation would include terms which represent the energy-momentum tensor of the sources. In the above equation $h = \eta^{\mu\nu} h_{\mu\nu} = h^\mu_\mu$, and indices on $h_{\mu\nu}$ can be raised or lowered using the Minkowski metric $\eta_{\mu\nu}$ and its inverse $\eta^{\mu\nu}$. Since every term in (10.89) contains two derivatives, this suggests that the fluctuations $h_{\mu\nu}$ are associated with massless excitations.

As we will see shortly, the equation of motion (10.89) is invariant under the gauge transformations discussed in Section 3.6:

$$\delta_0 h^{\mu\nu}(p) = i p^\mu \epsilon^\nu(p) + i p^\nu \epsilon^\mu(p)\,. \tag{10.90}$$

The infinitesimal gauge parameter $\epsilon^\mu(p)$ is a vector. In gravitation, the gauge invariance is reparameterization invariance: the choice of coordinate system used to parameterize spacetime does not affect the physics.

Let us verify that (10.89) is invariant under the gauge transformation (10.90). First we compute $\delta_0 h$ and find that

$$\delta_0 h = \eta_{\mu\nu} \delta_0 h^{\mu\nu} = i \eta_{\mu\nu} (p^\mu \epsilon^\nu + p^\nu \epsilon^\mu) = 2 i p \cdot \epsilon\,. \tag{10.91}$$

The resulting variation in $S^{\mu\nu}$ is therefore given by

$$\begin{aligned} \delta_0 S^{\mu\nu} = {} & i p^2 (p^\mu \epsilon^\nu + p^\nu \epsilon^\mu) - i p_\alpha p^\mu (p^\nu \epsilon^\alpha + p^\alpha \epsilon^\nu) \\ & - i p_\alpha p^\nu (p^\mu \epsilon^\alpha + p^\alpha \epsilon^\mu) + 2 i p^\mu p^\nu p \cdot \epsilon\,. \end{aligned} \tag{10.92}$$

But we can rewrite

$$\begin{aligned} \delta_0 S^{\mu\nu} = {} & i p^2 (p^\mu \epsilon^\nu + p^\nu \epsilon^\mu) - i p^\mu p^\nu (p \cdot \epsilon) - i p^2 p^\mu \epsilon^\nu \\ & - i p^\mu p^\nu (p \cdot \epsilon) - i p^2 p^\nu \epsilon^\mu + 2 i p^\mu p^\nu p \cdot \epsilon\,. \end{aligned} \tag{10.93}$$

It is readily seen that all the terms in (10.93) cancel, so $\delta_0 S^{\mu\nu} = 0$. The equation of motion exhibits the claimed gauge invariance.

Since the metric $h^{\mu\nu}$ is symmetric and has two indices, each running over $(+, -, I)$, we have the following objects to consider:

$$(h^{IJ}, h^{+I}, h^{-I}, h^{+-}, h^{++}, h^{--})\,. \tag{10.94}$$

We shall try to set to zero all the fields in (10.94) that contain a $+$ index. For this, we use (10.90) to examine their gauge transformations:

$$\delta_0 h^{++} = 2 i p^+ \epsilon^+\,, \tag{10.95}$$

$$\delta_0 h^{+-} = i p^+ \epsilon^- + i p^- \epsilon^+\,, \tag{10.96}$$

$$\delta_0 h^{+I} = i p^+ \epsilon^I + i p^I \epsilon^+\,. \tag{10.97}$$

As before, we assume $p^+ \neq 0$. From (10.95), we see that a judicious choice of ϵ^+ will permit us to gauge away h^{++}. This fixes our choice of ϵ^+. Looking at equation (10.96), we see that although we have fixed ϵ^+, we can still find an ϵ^- that will gauge away h^{+-}. This fixes ϵ^-. Similarly, we can use (10.97) and a suitable choice of ϵ^I to set h^{+I} to zero. We have used the full gauge freedom to set to zero all of the entries in $h^{\mu\nu}$ with $+$ indices. This defines the light-cone gauge for the gravity field:

$$\text{light-cone gauge conditions:} \quad h^{++} = h^{+-} = h^{+I} = 0. \tag{10.98}$$

The remaining degrees of freedom are carried by

$$(h^{IJ}, h^{-I}, h^{--}) \,. \tag{10.99}$$

We must now see what is implied by the equations of motion (10.89). Bearing in mind the gauge conditions (10.98), when $\mu = \nu = +$ we find

$$(p^+)^2 h = 0 \longrightarrow h = 0 \,. \tag{10.100}$$

More explicitly,

$$h = \eta_{\mu\nu} h^{\mu\nu} = -2h^{+-} + h^{II} = 0 \longrightarrow h^{II} = 0 \,, \tag{10.101}$$

since $h^{+-} = 0$ in our gauge. The equation $h^{II} = 0$ means that the matrix h^{IJ} is traceless. With $h = 0$, the equation of motion (10.89) reduces to

$$p^2 h^{\mu\nu} - p^\mu (p_\alpha h^{\nu\alpha}) - p^\nu (p_\alpha h^{\mu\alpha}) = 0 \,. \tag{10.102}$$

Now set $\mu = +$. We obtain $p^+ (p_\alpha h^{\nu\alpha}) = 0$, and as a result

$$p_\alpha h^{\nu\alpha} = 0 \,. \tag{10.103}$$

If (10.103) holds, equation (10.102) reduces to

$$p^2 h^{\mu\nu} = 0 \,. \tag{10.104}$$

This is all that remains of the equation of motion! Before delving into this familiar equation, let us investigate the implications of (10.103). The only free index here is ν. For $\nu = +$ the equation is trivial, since the $h^{+\alpha}$ are zero in our gauge. Consider now $\nu = I$. This gives $p_\alpha h^{I\alpha} = 0$, which we can expand as

$$-p^+ h^{I-} - p^- h^{I+} + p_J h^{IJ} = 0 \longrightarrow h^{I-} = \frac{1}{p^+} p_J h^{IJ} \,. \tag{10.105}$$

Similarly, with $\nu = -$ we get $p_\alpha h^{-\alpha} = 0$, which expands to

$$-p^+ h^{--} - p^- h^{-+} + p_I h^{-I} = 0 \longrightarrow h^{--} = \frac{1}{p^+} p_I h^{-I} \,. \tag{10.106}$$

Equations (10.105) and (10.106) give us the h with $-$ indices in terms of the transverse h^{IJ}. There is no additional content to (10.103).

We now return to equation (10.104). This equation holds trivially for any field with a + index. The equation is nontrivial for transverse indices:

$$p^2 h^{IJ}(p) = 0\,. \tag{10.107}$$

Equations $p^2 h^{I-} = 0$ and $p^2 h^{--} = 0$ are automatically satisfied on account of our solutions (10.105) and (10.106), together with (10.107). Equation (10.107) implies that $h^{IJ}(p) = 0$ for $p^2 \neq 0$, in which case all other components of $h^{\mu\nu}$ also vanish. For $p^2 = 0$, the $h^{IJ}(p)$ are unconstrained, except for the tracelessness condition $h_{II}(p) = 0$. All other components are determined in terms of the transverse components.

We conclude that the degrees of freedom of the classical D-dimensional gravitational field are carried by a *symmetric, traceless, transverse* tensor field h^{IJ}, the components of which satisfy the equation of motion of a massless scalar. This tensor has as many components as a symmetric traceless square matrix of size $(D-2)$. The number of components $n(D)$ in this matrix is

$$n(D) = \frac{1}{2}\,(D-2)(D-1) - 1 = \frac{1}{2}\,D(D-3)\,. \tag{10.108}$$

Moreover, as before, we count a massless scalar as one degree of freedom per point on the mass-shell. Therefore we say that a classical gravity wave has $n(D)$ degrees of freedom per point on the mass-shell. In four-dimensional spacetime there are two transverse directions, and a symmetric traceless 2×2 matrix has two independent components. In four dimensions we thus have $n(4) = 2$ degrees of freedom. In five dimensions we have $n(5) = 5$ degrees of freedom, in ten dimensions $n(10) = 35$ degrees of freedom, and in twenty-six dimensions $n(26) = 299$ degrees of freedom. To obtain graviton states, each of the independent classical fields h^{IJ} fields is expanded in terms of creation and annihilation operators, just as we did for the scalar field in (10.58). To do so we need oscillators $a^{IJ}_{p^+,\,p_T}$ and $a^{IJ\dagger}_{p^+,\,p_T}$. We introduce a vacuum $|\Omega\rangle$, and a basis of states

$$a^{IJ\,\dagger}_{p^+,\,p_T}|\Omega\rangle\,. \tag{10.109}$$

A one-graviton state with momentum $(p^+, \vec{p}_T)$ is a linear superposition of the above states:

$$\text{one-graviton states:} \quad \sum_{I,J=2}^{D-1} \xi_{IJ}\, a^{IJ\,\dagger}_{p^+,\,p_T}|\Omega\rangle\,, \quad \xi_{II} = 0. \tag{10.110}$$

Here ξ_{IJ} is the graviton polarization tensor. The classical tracelessness condition that we found earlier becomes in the quantum theory the tracelessness $\xi_{II} = 0$ of the polarization tensor. Since ξ_{IJ} is a traceless symmetric matrix of size $(D-2)$, we have $n(D)$ linearly independent graviton states for each point on the physical mass-shell.

Problems

Problem 10.1 Momentum for the classical scalar field.

Show that the integral (10.46), when evaluated for the field configuration (10.33) gives

$$\vec{P} = -\frac{i\,\vec{p}}{2E_p}\left(\dot{a}^* a - a^* \dot{a}\right) .$$

Use (10.44) to show that $\vec{P} = \vec{p}\left(a_p^* a_p - a_{-p}^* a_{-p}\right)$, as quoted in (10.47).

Problem 10.2 Commutator for the quantum scalar field.

(a) Consider a periodic function $f(\vec{x})$ on the box described above equation (10.34). Such a function can be expanded as the Fourier series

$$f(\vec{x}) = \sum_{\vec{p}} f(\vec{p})\, e^{i\,\vec{p}\cdot\vec{x}} . \tag{1}$$

Show that

$$f(\vec{p}) = \frac{1}{V}\int d\vec{x}'\; f(\vec{x}')e^{-i\,\vec{p}\cdot\vec{x}'} . \tag{2}$$

Plug (2) back into (1) to obtain an infinite sum representation for the d-dimensional delta function $\delta^d(\vec{x} - \vec{x}')$.

(b) Consider the complete scalar field expansion in (10.58). Calculate the corresponding expansion of $\Pi(t, \vec{x}) = \partial_0\phi(t, \vec{x})$. Show that

$$[\phi(t, \vec{x})\,,\; \Pi(t, \vec{x}')\,] = i\,\delta^d(\vec{x} - \vec{x}') . \tag{3}$$

This is the equal-time commutator between the field operator and its corresponding conjugate momentum. Most discussions of quantum field theory begin by postulating this commutator.

Problem 10.3 Light-cone components of Lorentz tensors.[†]

(a) Verify that the Lorentz covariant equation $A^\mu = B^\mu$, for $\mu = 0, 1, \ldots, d$, implies that $A^+ = B^+$, $A^- = B^-$, and $A^I = B^I$.

Given a Lorentz tensor $R^{\mu\nu}$, how do we define the light-cone components R^{+-}, R^{++}, ...? To find out, note that the definition must work for *any* tensor, so it must work when $R^{\mu\nu} = A^\mu B^\nu$. Thus, for example, $R^{+-} = A^+ B^-$, and writing A^+ and B^- in terms of Lorentz components, you can determine R^{+-} in terms of R^{00}, R^{01}, R^{10}, and R^{11}.

(b) Calculate R^{++}, R^{+-}, R^{-+}, and R^{--} in terms of the Lorentz components of $R^{\mu\nu}$. Explain why an equality $R^{\mu\nu} = S^{\mu\nu}$ between Lorentz tensors implies the equality of the light-cone components.

(c) Check that your result in (b) gives the expected answers for the light-cone components of the Minkowski metric: η^{++}, η^{+-}, η^{-+}, and η^{--}.

(d) Consider the antisymmetric electromagnetic field strength $F^{\mu\nu}$ in four dimensions. Calculate the light-cone components F^{+-}, F^{+I}, F^{-I}, and F^{IJ} in terms of the Lorentz components of $F^{\mu\nu}$. Rewrite your answers in terms of $\vec{E}$ and $\vec{B}$ fields.

Problem 10.4 Constant electric field in light-cone gauge.

Find potentials that describe a uniform constant electric field $\vec{E} = E_0\vec{e}_x$ in the light-cone gauge ($A^+ = (A^0 + A^1)/\sqrt{2} = 0$). Write A^- and A^I in terms of the light-cone coordinates x^+, x^-, and x^I.

Problem 10.5 Gravitational fields that are pure gauge.

Following the discussion of Maxwell fields that are pure gauge, define gravitational fields that are pure gauge. Prove that any gravitational field $h_{\mu\nu}(p)$ which satisfies the equations of motion is pure gauge when $p^2 \neq 0$.

Problem 10.6 Field equations and particle states for the Kalb–Ramond field $B_{\mu\nu}$.[†]

Here we examine the field theory of a massless antisymmetric tensor gauge field $B_{\mu\nu} = -B_{\nu\mu}$. This gauge field is a tensor analog of the Maxwell gauge field A_μ. In Maxwell theory we defined the field strength $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. For $B_{\mu\nu}$ we define a field strength $H_{\mu\nu\rho}$:

$$H_{\mu\nu\rho} \equiv \partial_\mu B_{\nu\rho} + \partial_\nu B_{\rho\mu} + \partial_\rho B_{\mu\nu} \,.$$

(a) Show that $H_{\mu\nu\rho}$ is totally antisymmetric. Prove that $H_{\mu\nu\rho}$ is invariant under the *gauge transformations*

$$\delta B_{\mu\nu} = \partial_\mu \epsilon_\nu - \partial_\nu \epsilon_\mu \,.$$

(b) The above gauge transformations are peculiar: the gauge parameters themselves have a gauge invariance! Show that ϵ'_μ given as

$$\epsilon'_\mu = \epsilon_\mu + \partial_\mu \lambda \,,$$

generates the *same* gauge transformations as ϵ_μ.

(c) Use light-cone coordinates and momentum space to argue that $\epsilon^+(p)$ can be set to zero for a suitable choice of $\lambda(p)$. Thus, the effective gauge symmetry of the Kalb–Ramond field is generated by the gauge parameters $\epsilon^I(p)$ and $\epsilon^-(p)$.

(d) Consider the spacetime action principle

$$S \sim \int d^D x \left(-\frac{1}{6} H_{\mu\nu\rho} H^{\mu\nu\rho}\right).$$

Find the field equation for $B_{\mu\nu}$, and write it in momentum space.

(e) What are the suitable light-cone gauge conditions for $B^{\mu\nu}$? Bearing in mind the results of part (c), show that these gauge conditions can be implemented using the gauge invariance. Analyze the equations of motion, show that $p^2 B^{\mu\nu} = 0$, and find the components of $B^{\mu\nu}$ which represent truly independent degrees of freedom.

(f) Argue that the one-particle states of the Kalb–Ramond field are

$$\sum_{I,J=2}^{D-1} \zeta_{IJ}\, a^{IJ\dagger}_{p^+,\,p_T} |\Omega\rangle \,. \tag{1}$$

What kind of matrix is ζ_{IJ}?

Problem 10.7 Massive vector field.†

The purpose of this problem is to understand the massive version of Maxwell fields. We will see that in D-dimensional spacetime a massive vector field has $(D-1)$ degrees of freedom.

Consider the action $S = \int d^D x\, \mathcal{L}$ with

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} - \frac{1}{2} m^2 A_\mu A^\mu - \frac{1}{2} \partial_\mu \phi\, \partial^\mu \phi + m A^\mu\, \partial_\mu \phi \,.$$

The first term in $\mathcal{L}$ is the familiar one for the Maxwell field. The second looks like a mass term for the Maxwell field, but alone would not suffice. The extra terms show the real scalar field ϕ that, as we shall see, is *eaten* to give the gauge field a mass.

(a) Show that the Lagrangian $\mathcal{L}$ is invariant under the infinitesimal gauge transformation

$$\delta A_\mu = \partial_\mu \epsilon \,, \quad \delta\phi = \ldots,$$

where the dots denote an expression that you must determine. While the gauge field has the familiar Maxwell gauge transformation, it is unusual to have a real scalar field with a gauge transformation.

(b) Vary the action and write down the field equations for A^μ and for ϕ.

(c) Argue that the gauge transformations allow us to set $\phi = 0$. Since the field ϕ disappears from sight, we say it is *eaten*. What do the field equations in part (b) simplify into?

(d) Write the simplified equations in momentum space and show that for $p^2 \neq -m^2$ there are no nontrivial solutions, while for $p^2 = -m^2$ the solution implies that there are $D-1$ degrees of freedom. (It may be useful to use a Lorentz transformation to represent the vector p^μ which satisfies $p^2 = -m^2$ as a vector that has a component only in one direction.)

11 The relativistic quantum point particle

To prepare ourselves for quantizing the string, we study the light-cone gauge quantization of the relativistic point particle. We set up the quantum theory by requiring that the Heisenberg operators satisfy the classical equations of motion. We show that the quantum states of the relativistic point particle coincide with the one-particle states of the quantum scalar field. Moreover, the Schrödinger equation for the particle wavefunctions coincides with the classical scalar field equations. Finally, we set up light-cone gauge Lorentz generators.

11.1 Light-cone point particle

In this section we study the classical relativistic point particle using the light-cone gauge. This is, in fact, a much easier task than the one we faced in Chapter 9, where we examined the classical relativistic string in the light-cone gauge. Our present discussion will allow us to face the complications of quantization in the simpler context of the particle. Many of the ideas needed to quantize the string are also needed to quantize the point particle.

The action for the relativistic point particle was studied in Chapter 5. Let us begin our analysis with the expression given in equation (5.15), where an arbitrary parameter τ is used to parameterize the motion of the particle:

$$S = -m \int_{\tau_i}^{\tau_f} \sqrt{-\eta_{\mu\nu} \frac{dx^\mu}{d\tau} \frac{dx^\nu}{d\tau}}\, d\tau \,. \tag{11.1}$$

In writing the above action, we have set $c = 1$. We will also set $\hbar = 1$ when appropriate. Finally, the parameter τ will be dimensionless, just as it was chosen to be for the relativistic string starting in Chapter 9. We can simplify our notation by writing

$$\eta_{\mu\nu} \frac{dx^\mu}{d\tau} \frac{dx^\nu}{d\tau} = \eta_{\mu\nu}\, \dot{x}^\mu\, \dot{x}^\nu = \dot{x}^2 \,. \tag{11.2}$$

Thinking of τ as a time variable and of the $x^\mu(\tau)$ as coordinates, the action S defines a Lagrangian L as

$$S = \int_{\tau_i}^{\tau_f} L\, d\tau \,, \quad L = -m\sqrt{-\dot{x}^2} \,. \tag{11.3}$$

As usual, the momentum is obtained by differentiating the Lagrangian with respect to the velocity:

$$p_\mu = \frac{\partial L}{\partial \dot{x}^\mu} = \frac{m\dot{x}_\mu}{\sqrt{-\dot{x}^2}} \,. \tag{11.4}$$

The Euler–Lagrange equations which arise from the Lagrangian are

$$\frac{dp_\mu}{d\tau} = 0 \,. \tag{11.5}$$

All components of the momentum are constants of the motion. Given (11.4) one readily checks that the momentum components satisfy the constraint

$$p^2 + m^2 = 0 \,. \tag{11.6}$$

To define the light-cone gauge for the particle, we set the coordinate x^+ of the particle proportional to τ:

$$\text{light-cone gauge condition:} \quad x^+ = \frac{1}{m^2}\, p^+ \tau \,. \tag{11.7}$$

The factor of m^2 on the right-hand side is needed to get the units to work. Now consider the $+$ component of equation (11.4):

$$p^+ = \frac{m}{\sqrt{-\dot{x}^2}}\, \dot{x}^+ = \frac{1}{\sqrt{-\dot{x}^2}}\, \frac{p^+}{m} \,. \tag{11.8}$$

Cancelling the common factor of p^+, and squaring, we find

$$\dot{x}^2 = -\frac{1}{m^2} \,. \tag{11.9}$$

This result helps us simplify the expression (11.4) for the momentum:

$$p_\mu = m^2\, \dot{x}_\mu \,. \tag{11.10}$$

The appearance of m^2, as opposed to m, is due to our choice of unitless τ. The equation of motion (11.5) then gives

$$\ddot{x}_\mu = 0 \,. \tag{11.11}$$

Expanding the constraint (11.6) in light-cone components,

$$-2p^+p^- + p^I p^I + m^2 = 0 \;\longrightarrow\; p^- = \frac{1}{2p^+}(p^I p^I + m^2) \,. \tag{11.12}$$

The value of p^- is determined by p^+ and the components p^I of the transverse momentum $\vec{p}_T$. Having solved for p^-, equation (11.10) gives

$$\frac{dx^-}{d\tau} = \frac{1}{m^2}\, p^- \,, \tag{11.13}$$

which is integrated to find

$$x^-(\tau) = x_0^- + \frac{p^-}{m^2}\,\tau\,, \tag{11.14}$$

where x_0^- is a constant of integration. Equation (11.10) also gives $dx^I/d\tau = p^I/m^2$, which is integrated to give

$$x^I(\tau) = x_0^I + \frac{p^I}{m^2}\,\tau\,, \tag{11.15}$$

where x_0^I is a constant of integration. Note that the light-cone gauge condition (11.7) implies that $x^+(\tau)$ has no constant piece x_0^+.

The specification of the motion of the point particle is now complete. Equation (11.12) tells us that the momentum is completely determined once we fix p^+ and the components p^I of the transverse momentum $\vec{p}_T$. The motion in the x^- direction is determined by (11.14), once we fix the value of x_0^-. The transverse motion is determined by the $x^I(\tau)$, or the x_0^I, since we presume to know the p^I. For a symmetric treatment of coordinates versus momenta in the quantum theory, we choose the x^I as dynamical variables. Our independent dynamical variables for the point particle are therefore

$$\text{Dynamical variables:} \quad \left(x^I, \quad x_0^-, \quad p^I, \quad p^+\right). \tag{11.16}$$

11.2 Heisenberg and Schrödinger pictures

Traditionally, there are two main approaches to the understanding of time evolution in quantum mechanics. In the Schrödinger picture, the state of a system evolves in time, while operators remain unchanged. In the Heisenberg picture, it is the operators which evolve in time, while the state remains unchanged. Of the two, the Heisenberg picture is more closely related to classical mechanics, where the dynamical variables (which become operators in quantum mechanics) evolve in time. Both the Schrödinger and the Heisenberg pictures will be useful in developing the quantum theories of the relativistic point particle and the relativistic string. Because we would like to exploit our understanding of classical dynamics in developing the quantum theories, we will begin by focusing on the Heisenberg picture.

Both the Heisenberg and the Schrödinger picture make use of the same state space. Whereas in the Heisenberg picture the state which represents a particular physical system is fixed in time, in the Schrödinger picture the state of a system is constantly changing direction in the state space in a manner which is determined by the Schrödinger equation. Although we generally think of the operators in the Schrödinger picture as being time independent, there are those which depend *explicitly* on time and therefore have time dependence. These operators are formed from time-independent operators and the variable t. For example, the position and momentum operators q and p are time independent. But

the operator $\mathcal{O} = q + p\,t$ has explicit time dependence. If it has explicit time dependence, even the Hamiltonian $H(p, q; t)$ can be a time-dependent operator.

Now, as we move from the Schrödinger to the Heisenberg picture, we will encounter operators with two types of time dependence. As we noted earlier, Heisenberg operators have time dependence, but this time dependence can be both *implicit* and *explicit*. The Heisenberg equivalent of a time-independent Schrödinger operator is said to have implicit time dependence. This implicit time dependence is due to our folding into the operators the time dependence which, in the Schrödinger picture, is present in the state. If a Heisenberg operator is explicitly time dependent it is because the explicit time dependence of the corresponding Schrödinger operator has been carried over.

When we pass from the Schrödinger to the Heisenberg picture, the time-independent Schrödinger operators q and p, for example, become $q(t)$ and $p(t)$, respectively. The Schrödinger commutator $[q, p] = i$ turns into the commutator

$$[\,q(t)\,,\ p(t)\,] = i\,. \tag{11.17}$$

Although $q(t)$ and $p(t)$ depend on time, their time dependence is implicit. If $\xi(t)$ is a Heisenberg operator arising from a time-*independent* Schrödinger operator, the time evolution of $\xi(t)$ is governed by

$$i\frac{d\xi(t)}{dt} = \Big[\,\xi(t)\,,\ H(p(t), q(t); t)\Big]\,. \tag{11.18}$$

Here $H(p(t), q(t); t)$ is the Heisenberg Hamiltonian corresponding to the possibly time-dependent Schrödinger Hamiltonian $H(p, q; t)$.

If $\mathcal{O}(t)$ is the Heisenberg operator which corresponds to an explicitly time-*dependent* Schrödinger operator, then the time evolution of $\mathcal{O}(t)$ is given by

$$i\,\frac{d\mathcal{O}(t)}{dt} = i\,\frac{\partial\mathcal{O}}{\partial t} + \Big[\mathcal{O}(t)\,,\ H(p(t), q(t); t)\Big]\,. \tag{11.19}$$

This equation reduces to (11.18) when the operator has no explicit time dependence. If the Hamiltonian $H(p(t), q(t))$ has no explicit time dependence, then we can use (11.18) with $\xi = H$, to find

$$\frac{d}{dt}H(p(t), q(t)) = 0\,. \tag{11.20}$$

In this case the Hamiltonian is a constant of the motion.

The discussion above must be supplemented by rules to pass from Schrödinger to Heisenberg operators. Assume that the Schrödinger Hamiltonian $H(p, q)$ is time independent. In this case a state $|\Psi\rangle$ at time $t = 0$ evolves in time, and is given at time t by

$$|\Psi, t\rangle = e^{-iHt}|\Psi\rangle\,. \tag{11.21}$$

Quick calculation 11.1 Confirm that $|\Psi, t\rangle$ satisfies the Schrödinger equation

$$i\frac{d}{dt}|\Psi, t\rangle = H|\Psi, t\rangle\,. \tag{11.22}$$

It is clear from (11.21) that the operator e^{iHt} brings time-dependent states to rest:

$$e^{iHt}|\Psi, t\rangle = |\Psi\rangle\,. \tag{11.23}$$

If we act with this operator on the product $\alpha|\Psi, t\rangle$, where α is a Schrödinger operator, we find

$$e^{iHt}\alpha|\Psi, t\rangle = e^{iHt}\,\alpha\, e^{-iHt}|\Psi\rangle \equiv \alpha(t)|\Psi\rangle\,, \tag{11.24}$$

where $\alpha(t) = e^{iHt}\alpha\, e^{-iHt}$ is the Heisenberg operator corresponding to the Schrödinger operator α. This definition also applies for a Schrödinger operator $\alpha_{\rm S}(t)$ that has explicit time dependence, in which case the corresponding Heisenberg operator is $\alpha_{\rm H}(t) = e^{iHt}\alpha_{\rm S}(t)\, e^{-iHt}$. The construction ensures that given a set of Schrödinger operators that satisfy certain commutation relations, the corresponding Heisenberg operators satisfy the same commutation relations.

Quick calculation 11.2 If $[\alpha_1, \alpha_2] = \alpha_3$ holds for Schrödinger operators α_1, α_2, and α_3, show that $[\alpha_1(t), \alpha_2(t)] = \alpha_3(t)$ holds for the corresponding Heisenberg operators.

This result holds even if the Hamiltonian is time dependent (Problem 11.2). It justifies the commutator in (11.17), noting that the constant right-hand side is not affected by the rule turning a Schrödinger operator into a Heisenberg operator.

11.3 Quantization of the point particle

We now develop a quantum theory from the classical theory of the relativistic point particle. We will define the relevant Schrödinger and Heisenberg operators, including the Hamiltonian, and describe the state space. All of this will be done in the light-cone gauge.

Our first step is to choose a set of time-independent Schrödinger operators. A reasonable choice is provided by the dynamical variables in (11.16):

$$\text{time-independent Schrödinger operators:}\quad \left(x^I,\quad x_0^-,\quad p^I,\quad p^+\right). \tag{11.25}$$

We could include hats to distinguish the operators from their eigenvalues, but this will not be necessary in most cases. We parameterize the trajectory of a point particle using τ, so the associated Heisenberg operators are:

$$\text{Heisenberg operators:} \quad \Big(x^I(\tau), \quad x_0^-(\tau), \quad p^I(\tau), \quad p^+(\tau) \Big) . \tag{11.26}$$

We postulate the following commutation relations for the Schrödinger operators:

$$[\, x^I, \, p^J] = i \, \eta^{IJ} , \quad [\, x_0^-, \, p^+] = i \, \eta^{-+} = -i , \tag{11.27}$$

with all other commutators set equal to zero. The first commutator is the familiar commutator of spatial coordinates with the corresponding spatial momenta (recall that $\eta^{IJ} = \delta^{IJ}$). The second commutator is well motivated, after all, x_0^- is treated as a spatial coordinate in the light-cone, and p^+ is the corresponding conjugate momentum. The second commutator, just like the first one, has an η carrying the indices of the coordinate and the momentum.

The Heisenberg operators, as explained earlier, satisfy the same commutation relations as the Schrödinger operators:

$$[\, x^I(\tau), \, p^J(\tau)] = i \, \eta^{IJ} , \quad [\, x_0^-(\tau), \, p^+(\tau)] = -i , \tag{11.28}$$

with all other commutators set equal to zero.

We have discussed the operators that correspond to the independent observables of the classical theory. But just as there are classical observables which depend on those independent ones, there are also quantum operators which are constructed from the set of independent Schrödinger operators, and time. These additional operators are $x^+(\tau), x^-(\tau)$ and p^-. These operators are *defined* using the quantum analogs of equations (11.7), (11.14), and (11.12):

$$x^+(\tau) \equiv \frac{p^+}{m^2} \tau , \tag{11.29}$$

$$x^-(\tau) \equiv x_0^- + \frac{p^-}{m^2} \tau , \tag{11.30}$$

$$p^- \equiv \frac{1}{2p^+} \Big(p^I p^I + m^2 \Big) . \tag{11.31}$$

Note that p^- is time independent. Both $x^+(\tau)$ and $x^-(\tau)$ are time-dependent Schrödinger operators.

The commutation relations involving the operators $x^+(\tau), x^-(\tau)$, and p^- are determined by the postulated commutation relations in (11.27), along with the defining equations (11.29)–(11.31). The decision to choose the operators in (11.25) as the independent operators of our quantum theory was very significant. For example, if we had chosen x^+ and p^- to be independent operators, we might have been led to write a commutation relation $[x^+, p^-] = -i$. In our present framework, however, this quantity vanishes, since $[\, p^+, p^I] = 0$.

We have not yet determined the Hamiltonian H. Since p^- is the light-cone energy (see (2.94)), we expect it to generate x^+ evolution:

$$\frac{\partial}{\partial x^+} \longleftrightarrow p^- . \tag{11.32}$$

Although x^+ is light-cone time, we are parameterizing our operators with τ, so we expect H to generate τ evolution, which is related, but is not the same as x^+ evolution. Since $x^+ = p^+\tau/m^2$, we can anticipate that τ evolution will be generated by

$$\frac{\partial}{\partial \tau} = \frac{p^+}{m^2}\frac{\partial}{\partial x^+} \longleftrightarrow \frac{p^+}{m^2}\, p^- . \tag{11.33}$$

We therefore postulate the Heisenberg Hamiltonian

$$H(\tau) = \frac{p^+(\tau)}{m^2}\; p^-(\tau) = \frac{1}{2m^2}\Big(p^I(\tau)p^I(\tau) + m^2\Big) . \tag{11.34}$$

Note that $H(\tau)$ has no explicit time dependence. Equation (11.20) applies, and as a result, the Hamiltonian is actually time independent.

Let us now make sure that this Hamiltonian generates the expected equations of motion. First we check that H gives the correct time evolution of the Heisenberg operators (11.26) which arise from the time-independent Schrödinger operators. The equation governing the time evolution of those operators is (11.18). Let us begin with p^+ and p^I:

$$\begin{aligned} i\,\frac{dp^+(\tau)}{d\tau} &= \big[\, p^+(\tau)\,, H(\tau)\,\big] = 0, \\ i\,\frac{dp^I(\tau)}{d\tau} &= \Big[\, p^I(\tau)\,, H(\tau)\,\Big] = 0. \end{aligned} \tag{11.35}$$

Both of these commutators vanish because H is a function of $p^I(\tau)$ alone, and all the momenta commute. Equations (11.35) are good news, because the classical momenta p^+ and p^I are constants of the motion. This allows us to write $p^I(\tau) = p^I$ and $p^+(\tau) = p^+$. We now test the τ development of the Heisenberg operator $x^I(\tau)$:

$$i\,\frac{dx^I(\tau)}{d\tau} = \Big[\, x^I(\tau)\,,\; \frac{1}{2m^2}\Big(p^J p^J + m^2\Big)\Big] = i\,\frac{p^I}{m^2} . \tag{11.36}$$

Here, we have used $[x^I, p^J p^J] = [x^I, p^J]p^J + p^J[x^I, p^J] = 2i\, p^I$. Cancelling the common factor of i in (11.36), we find

$$\frac{dx^I(\tau)}{d\tau} = \frac{p^I}{m^2} . \tag{11.37}$$

This result is in accord with our classical expectations and allows us to write

$$x^I(\tau) = x_0^I + \frac{p^I}{m^2}\,\tau\,, \tag{11.38}$$

where x_0^I is an operator without any time dependence. Finally, we must examine $x_0^-(\tau)$. Since $x_0^-(\tau)$ commutes with $p^I(\tau)$,

$$i\,\frac{dx_0^-(\tau)}{d\tau} = \Big[\, x_0^-(\tau)\,,\ \frac{1}{2m^2}\Big(p^I p^I + m^2\Big)\Big] = 0\,. \tag{11.39}$$

As expected, this operator is a constant of the motion, and we can write $x_0^-(\tau) = x_0^-$. So as far as the operators in (11.25) are concerned, our ansatz for H functions properly as a Hamiltonian.

We now turn to the remaining operators $x^+(\tau)$, $x^-(\tau)$, and $p^-(\tau)$. Of these, $p^-(\tau)$ is a function of the p^I only and is therefore time independent. It is easy to to see that the commutator with H vanishes, so we have nothing left to check for this operator. The Heisenberg operators $x^+(\tau)$ and $x^-(\tau)$ both arise from Schrödinger operators with explicit time dependence, so we use (11.19) to calculate their time evolution. For example:

$$i\frac{dx^-(\tau)}{d\tau} = i\frac{\partial x^-}{\partial \tau} + \big[\, x^-(\tau)\,,\ H(\tau)\,\big]\,. \tag{11.40}$$

Since $x^-(\tau) \equiv x_0^- + p^-\tau/m^2$ and both x_0^- and p^- commute with the p^I, we see that $[x^-(\tau), H(\tau)] = 0$. Consequently,

$$\frac{dx^-(\tau)}{d\tau} = \frac{p^-}{m^2}\,, \tag{11.41}$$

which is the expected result. Similarly, since $x^+(\tau) = p^+\tau/m^2$, we find that $[x^+(\tau), H(\tau)] = 0$, and therefore

$$\frac{dx^+(\tau)}{d\tau} = \frac{\partial x^+}{\partial \tau} = \frac{p^+}{m^2}\,. \tag{11.42}$$

These computations show that our ansatz (11.34) for the Hamiltonian generates the expected equations of operator evolution.

Quick calculation 11.3 We introduced x_0^I in (11.38) as a constant operator. Show that $dx_0^I/d\tau$ must be calculated by viewing x_0^I as the explicitly time-dependent Heisenberg operator defined by (11.38).

To complete our construction of the point particle quantum theory we must develop the state space, set up the Schrödinger equation, and define physical states. The time-independent states of the quantum theory are labeled by the eigenvalues of a maximal set of commuting operators. For the set of operators introduced in (11.25), a maximal commuting subset can include only one element from the pair $(x^-,\, p^+)$, and one element from each of the pairs $(x^I,\, p^I)$. Because it is usually convenient to work in momentum space, we will work with the operators p^+ and p^I. So we write the states as

$$\text{states of the quantum point particle:} \quad |p^+,\ \vec{p}_T\rangle\,, \tag{11.43}$$

where p^+ is the eigenvalue of the p^+ operator, and $\vec{p}_T$ is the transverse momentum, the components of which are the eigenvalues of the p^I operators:

$$\hat{p}^+|p^+,\ \vec{p}_T\rangle = p^+|p^+,\ \vec{p}_T\rangle\,, \quad \hat{p}^I|p^+,\ \vec{p}_T\rangle = p^I|p^+,\ \vec{p}_T\rangle\,. \tag{11.44}$$

In light of (11.31), these equations imply

$$\hat{p}^-|p^+,\ \vec{p}_T\rangle = \frac{1}{2p^+}\Big(p^I p^I + m^2\Big)|p^+,\ \vec{p}_T\rangle\,. \tag{11.45}$$

In addition, the Hamiltonian (11.34) acting on the states gives

$$H\,|p^+,\ \vec{p}_T\rangle = \frac{1}{2m^2}\Big(p^I p^I + m^2\Big)|p^+,\ \vec{p}_T\rangle\,. \tag{11.46}$$

It then follows that the time-dependent states

$$\exp\Big(-i\,\frac{1}{2m^2}\big(p^I p^I + m^2\big)\,\tau\Big)\ |p^+,\ \vec{p}_T\rangle \tag{11.47}$$

satisfy the Schrödinger equation. They are the *physical* time-dependent states associated with the states in (11.43).

More generally, we consider time-dependent superpositions of the basis states in (11.43):

$$|\Psi,\tau\rangle = \int dp^+ d\vec{p}_T\ \psi(\tau,\,p^+,\,\vec{p}_T)\,|p^+,\,\vec{p}_T\rangle\,. \tag{11.48}$$

Since p^+ and $\vec{p}_T$ are continuous variables, an integral is necessary. To produce a general τ-dependent superposition, we introduced the arbitrary function $\psi(\tau,\,p^+,\,\vec{p}_T)$. In fact, this function is the momentum-space wavefunction associated with the state $|\Psi,\tau\rangle$. Indeed, with dual bras $\langle p^+,\,\vec{p}_T|$ defined to satisfy

$$\langle p'^+,\,\vec{p}\,'_T|\,p^+,\,\vec{p}_T\rangle = \delta(p'^+ - p^+)\,\delta\,(\vec{p}\,'_T - \vec{p}_T)\,, \tag{11.49}$$

we see that

$$\langle p^+,\,\vec{p}_T|\,\Psi,\tau\rangle = \psi(\tau,\,p^+,\,\vec{p}_T)\,. \tag{11.50}$$

The Schrödinger equation for the state $|\Psi,\tau\rangle$ is

$$i\,\frac{\partial}{\partial\tau}|\Psi,\tau\rangle = H|\Psi,\tau\rangle\,. \tag{11.51}$$

Using the state in (11.48) and the Hamiltonian in (11.34), we find

$$\int dp^+ d\vec{p}_T\,\Big[i\,\frac{\partial}{\partial\tau}\psi(\tau,\,p^+,\,\vec{p}_T) - \frac{1}{2m^2}\Big(p^I p^I + m^2\Big)\psi(\tau,\,p^+,\,\vec{p}_T)\Big]|p^+,\,\vec{p}_T\rangle = 0\,. \tag{11.52}$$

Since the basis vectors $|p^+,\,\vec{p}_T\rangle$ are all linearly independent, the expression within brackets must vanish for all values of the momenta:

$$i\,\frac{\partial}{\partial\tau}\,\psi\,(\tau,\,p^+,\,\vec{p}_T) = \frac{1}{2m^2}\Big(p^I p^I + m^2\Big)\psi(\tau,\,p^+,\,\vec{p}_T)\,. \tag{11.53}$$

We recognize this equation as a Schrödinger equation for the momentum-space wavefunction $\psi(\tau,\,p^+,\,\vec{p}_T)$. If the wavefunction satisfies the Schrödinger equation, the state $|\Psi,\tau\rangle$ is a physical time-dependent state. This completes our development of the point particle quantum theory.

11.4 Quantum particle and scalar particles

The states of the quantum point particle given in (11.43) may remind you of the one-particle states (10.66) in the quantum theory of the scalar field. This is actually a fundamental correspondence.

There is a natural identification of the quantum states of a relativistic point particle of mass m with the one-particle states of the quantum theory of a scalar field of mass m:

$$|p^+,\ \vec{p}_T\rangle \longleftrightarrow a^\dagger_{p^+,\,p_T}|\Omega\rangle. \tag{11.54}$$

The identification is possible because the labels of the point particle states match with the labels of the creation operators which generate the one-particle states of the scalar quantum field theory. The correspondence between the quantum point particle and the quantum scalar field theory can be extended from the state space to the operators that act on the state space. The quantum point particle theory has operators p^+, p^I, and p^-, and so does the quantum field theory, as shown in (10.67). If we identify the state spaces using (11.54), then the two sets of operators give the same eigenvalues. This makes the identification natural.

The above observations lead us to conclude that the states of the quantum point particle and the one-particle states of the scalar field theory are indistinguishable. Because it contains creation operators that can act multiple times on the vacuum state, the scalar field theory has multiparticle states that did not arise in our quantization of the point particle. Indeed, there are no creation operators in the theory of the quantum point particle. Because it provides a natural description of multiparticle states, the scalar field theory can be said to be a more complete theory.

How could we have anticipated that the one-particle states of a quantum *scalar* field theory would match those of the quantum point particle? The answer is quite interesting: the Schrödinger equation for the quantum point particle wavefunctions has the form of the classical field equation for the scalar field. More precisely:

There is a canonical correspondence between the quantum point particle wavefunctions and the classical scalar field, such that the Schrödinger equation for the quantum point particle wavefunctions becomes the classical field equation for the scalar field.

One element of this correspondence is the classical field equation for the scalar field. In light-cone gauge, this equation takes the form (10.31):

$$\left(i\,\frac{\partial}{\partial\tau} - \frac{1}{2m^2}(p^I p^I + m^2)\right)\phi(\tau,\, p^+,\, \vec{p}_T) = 0\,. \tag{11.55}$$

This differential equation is first order in τ. The other element in the correspondence is this Schrödinger equation (11.53). The two equations are identical once we identify the wavefunction $\psi(\tau, p^+, \vec{p}_T)$ and the scalar field $\phi(\tau, p^+, \vec{p}_T)$:

$$\psi(\tau, p^+, \vec{p}_T) \longleftrightarrow \phi(\tau, p^+, \vec{p}_T)\,. \tag{11.56}$$

This is the claimed correspondence.

The quantization of the point particle is an example of *first quantization*. In first quantization, the coordinates and momenta of classical mechanics are turned into quantum operators and a state space is constructed. Generically, the result is a set of one-particle states. *Second quantization* refers to the quantization of a classical field theory, the result of which is a quantum field theory with field operators and multiparticle states. Our analysis allows us to see how second quantization follows after first quantization. A first quantization of the classical point particle mechanics gives one-particle states. We then reinterpret the Schrödinger equation for the associated wavefunctions as the classical field equation for a scalar field. A second quantization, this time of the classical field theory, gives us the set of multiparticle states.

So far we have only quantized the *free* relativistic point particle. All quantum states, including the multiparticle ones obtained by second quantization, represent free particles. How do we get interactions between the particles? Such processes are included in the scalar field theory by adding interaction terms to the action. All the terms that we have included so far are quadratic in the fields. The interaction terms include three or more fields. Since the quantum point particle state space does not include multiparticle states, the description of interactions in the language of first quantization is not straightforward. On the other hand, in the framework of quantum field theory interactions are dealt with very naturally.

11.5 Light-cone momentum operators

Since the point particle Lagrangian L in (11.3) depends only on τ derivatives of the coordinates, it is invariant under the translations

$$\delta x^\mu(\tau) = \epsilon^\mu\,, \tag{11.57}$$

with ϵ^μ constant. The conserved charge associated with this symmetry transformation is the momentum p_μ of the particle. This follows from (8.16) and (11.4).

What happens to conserved charges in the quantum theory? They become quantum operators with a remarkable property: they generate, via commutation, a quantum version of the symmetry transformation that gave rise to them classically!

This property is most apparent if we use a framework where the manifest Lorentz invariance of the classical theory is preserved in the quantization. This is *not* the framework we

have used to quantize the point particle. In light-cone gauge quantization, the x^0 and x^1 coordinates of the particle are afforded special treatment. This hides the Lorentz invariance of the theory from plain view. We will not discuss fully the Lorentz covariant quantization of the point particle. A few remarks will suffice for our present purposes. The covariant quantization of the string is discussed in some detail in Chapter 24.

In the Lorentz covariant quantization of the point particle, we have Heisenberg operators $x^\mu(\tau)$ and $p^\mu(\tau)$. Note that even the time coordinate $x^0(\tau)$ becomes an operator! The commutation relations are

$$[\, x^\mu(\tau),\, p^\nu(\tau)] = i\,\eta^{\mu\nu}\,, \tag{11.58}$$

as well as

$$[\, x^\mu(\tau),\, x^\nu(\tau)] = 0 \quad \text{and} \quad [\, p^\mu(\tau),\, p^\nu(\tau)] = 0\,. \tag{11.59}$$

Equation (11.58) is reasonable. The indices match, which ensures consistency with Lorentz covariance. Moreover, when μ and ν take spatial values, the commutation relations are the familiar ones. We already know that (11.58) is not consistent with the light-cone gauge commutators of Section 11.3. We saw there that $[x^+(\tau),\, p^-(\tau)] = 0$, while (11.58) would predict a nonzero result. An equality of two objects which carry Lorentz indices applies when the indices run over the light-cone values $+$, $-$, and I. The equation $R^{\mu\nu} = S^{\mu\nu}$, gives, for example, $R^{+-} = S^{+-}$ (Problem 10.3). As a result, equation (11.58) indeed gives $[x^+(\tau),\, p^-(\tau)] = i\eta^{+-} = -i$. Let us now check that the operator $p^\mu(\tau)$ generates translations. More precisely, we check that $i\epsilon_\rho\, p^\rho(\tau)$ generates the translation (11.57):

$$\delta x^\mu(\tau) = \big[\, i\epsilon_\rho\, p^\rho(\tau)\,,\, x^\mu(\tau)\,\big] = i\epsilon_\rho\,(-i\eta^{\rho\mu}) = \epsilon^\mu\,. \tag{11.60}$$

This is an elegant result, but it is by no means clear that it carries over to our light-cone gauge quantization. We must find out whether the light-cone gauge momentum operators generate translations.

For this purpose, we expand the generator $i\epsilon_\rho\, p^\rho(\tau)$ in light-cone components:

$$i\epsilon_\rho\, p^\rho(\tau) = -i\epsilon^-\, p^+ - i\epsilon^+\, p^- + i\epsilon^I\, p^I\,. \tag{11.61}$$

We have dropped the τ arguments from the momenta because they are τ-independent. Note that here, p^- is given by (11.31). Let us test (11.60) with $\epsilon^I \neq 0$, and $\epsilon^+ = \epsilon^- = 0$:

$$\delta x^\mu(\tau) = i\epsilon^I[\, p^I\,,\, x^\mu(\tau)\,]\,. \tag{11.62}$$

We would expect that $\delta x^J(\tau) = \epsilon^J$ and that $\delta x^+(\tau) = \delta x^-(\tau) = 0$. All these expectations are realized. Choosing $\mu = J$, and using the commutator (11.28) we find $\delta x^J(\tau) = \epsilon^J$. To compute the action on $x^+(\tau)$ and $x^-(\tau)$, we must use their definitions:

$$x^+(\tau) = \frac{p^+}{m^2}\tau\,, \quad x^-(\tau) = x_0^- + \frac{p^-}{m^2}\tau\,. \tag{11.63}$$

Recalling that p^I commutes with all momenta and with x_0^-, we confirm that $\delta x^+(\tau) = \delta x^-(\tau) = 0$.

Quick calculation 11.4 Test (11.60) with $\epsilon^- \neq 0$ and $\epsilon^+ = \epsilon^I = 0$. To do this compute $\delta x^\mu(\tau) = -i\epsilon^-[\, p^+\,,\, x^\mu(\tau)\,]$. Confirm that $\delta x^-(\tau) = \epsilon^-$ and that all other coordinates are not changed.

It remains to see whether p^- generates the expected translations. Since p^- is a nontrivial function of other momenta, there is some scope for complications! This time we consider the transformations that are generated using (11.60) with $\epsilon^+ \neq 0$ and $\epsilon^- = \epsilon^I = 0$:

$$\delta x^\mu(\tau) = -i\epsilon^+ [\, p^- , x^\mu(\tau)\,] . \tag{11.64}$$

The naive expectation $\delta x^+(\tau) = \epsilon^+$ is not realized: choosing $\mu = +$ and using (11.63) we see that

$$\delta x^+(\tau) = -i\epsilon^+ \Big[\, p^- , p^+ \frac{\tau}{m^2}\Big] = 0 . \tag{11.65}$$

Not only is $x^+(\tau)$ left unchanged, but the other components, which naively should be left unchanged, are not:

$$\delta x^I(\tau) = -i\epsilon^+ \Big[\, p^-, x^I(\tau)\Big] = -i\epsilon^+ \frac{1}{2p^+}\,(-2ip^I) \;= -\epsilon^+ \frac{p^I}{p^+} , \tag{11.66}$$

$$\delta x^-(\tau) = -i\epsilon^+ \Big[\, p^-, x_0^- + \frac{p^-}{m^2}\tau\Big] = -i\epsilon^+ [\, p^-, x_0^-] = -\epsilon^+ \frac{p^-}{p^+} . \tag{11.67}$$

In these calculations only one step requires some explanation. How do we find $[\, p^-, x_0^-]$? The only reason p^- does not commute with x_0^- is that p^- depends on p^+. In fact, what we need to know is the commutator $[x_0^-, 1/p^+]$. This can be found as follows:

$$\begin{aligned}\Big[\, x_0^- , \; \frac{1}{p^+}\,\Big] &= x_0^- \frac{1}{p^+} - \frac{1}{p^+} x_0^- = \frac{1}{p^+}\, p^+ x_0^- \frac{1}{p^+} - \frac{1}{p^+} x_0^- p^+ \frac{1}{p^+} \\ &= \frac{1}{p^+}\,[\, p^+ , x_0^-\,] \frac{1}{p^+} = \frac{i}{p^{+2}} .\end{aligned} \tag{11.68}$$

Quick calculation 11.5 Use (11.68) to show that

$$[\, x_0^- \,,\; p^-\,] = i\,\frac{p^-}{p^+} . \tag{11.69}$$

Equations (11.65), (11.66), and (11.67) show that p^- does not generate the expected transformations. What happened? It turns out that p^- actually generates both a translation *and* a reparameterization of the world-line of the particle. We know that the particle action is invariant under changes of parameterization $\tau \to \tau'(\tau)$. When we described symmetries in Chapter 8, however, we exhibited them as changes in the dynamical variables of the system. A change in parameterization can also be described in that way. Writing $\tau \to \tau' = \tau + \lambda(\tau)$, with λ infinitesimal, we note that the plausible change

$$x^\mu(\tau) \;\longrightarrow\; x^\mu(\tau + \lambda(\tau)) = x^\mu(\tau) + \lambda(\tau)\partial_\tau x^\mu(\tau) , \tag{11.70}$$

leads us to write

$$\delta x^\mu(\tau) = \lambda(\tau)\partial_\tau x^\mu(\tau) . \tag{11.71}$$

We claim that these are *symmetries* of the point particle theory. Actually, the variation (11.71) does not leave the point particle Lagrangian invariant. The Lagrangian changes into a total τ derivative (Problem 11.4), but this, in fact, suffices to have a symmetry (see Problem 8.9).

Let us now show that p^- generates a translation plus a reparameterization. The expected translation was $\delta x^+ = \epsilon^+$. On the other hand, from (11.71), a reparameterization of x^+ gives $\delta x^+ = \lambda \partial_\tau x^+$. Bearing in mind (11.65), the expected translation plus the reparameterization give zero variation, so,

$$0 = \epsilon^+ + \lambda \partial_\tau x^+(\tau) = \epsilon^+ + \lambda \frac{p^+}{m^2} \quad \longrightarrow \quad \lambda = -\frac{m^2}{p^+}\epsilon^+ . \tag{11.72}$$

The reparameterization parameter λ turns out to be a constant. We can now use this result to "explain" the transformations (11.66) and (11.67) that p^- generates on x^I and on x^-. For these coordinates there is no translation, but the reparameterization still applies. Therefore,

$$\delta x^I(\tau) = \lambda\, \partial_\tau x^I(\tau) = -\frac{m^2}{p^+}\,\epsilon^+ \frac{p^I}{m^2} = -\epsilon^+ \frac{p^I}{p^+} , \tag{11.73}$$

$$\delta x^-(\tau) = \lambda\, \partial_\tau x^-(\tau) = -\frac{m^2}{p^+}\,\epsilon^+ \frac{p^-}{m^2} = -\epsilon^+ \frac{p^-}{p^+} , \tag{11.74}$$

in perfect agreement with the transformations generated by p^-. We can also understand why p^- does not change x^+. If x^+ had been changed by a constant ϵ^+, the new x^+ coordinate would not satisfy the light-cone gauge condition whereby x^+ is just proportional to τ. In fact, p^- generates a translation plus the compensating transformation needed to preserve the light-cone gauge condition! That transformation turned out to be a reparameterization of the world-line.

One final remark about momentum operators. The Lorentz covariant momentum operators that we used to motivate our analysis generate simple translations and commute among each other. It follows directly that, using light-cone *coordinates*, the operators $p^\pm = (p^0 \pm p^1)/\sqrt{2}$ and the transverse p^I all commute. The light-cone *gauge* momentum operators we discussed above are completely different objects. They had an intricate action on coordinates, and p^- was defined in terms of the transverse momenta and p^+. Nevertheless, all the light-cone gauge momentum operators still commute. They obey the same commutation relations that the covariant operators do when expressed using light-cone coordinates.

11.6 Light-cone Lorentz generators

In Section 8.5 we determined the conserved charges that are associated with the Lorentz invariance of the relativistic string Lagrangian. Similar charges exist for the relativistic point particle. As we found in (8.52), the infinitesimal Lorentz transformations of the point particle coordinates $x^\mu(\tau)$ take the form

$$\delta x^\mu(\tau) = \epsilon^{\mu\nu} x_\nu(\tau) , \tag{11.75}$$

where $\epsilon^{\mu\nu} = -\epsilon^{\nu\mu}$ are a set of infinitesimal constants. The associated Lorentz charges are given by

$$M^{\mu\nu} = x^\mu(\tau)p^\nu(\tau) - x^\nu(\tau)p^\mu(\tau)\,, \tag{11.76}$$

as you may have derived in Problem 8.5. These charges are conserved classically. The quantum charges are expected to generate Lorentz transformations of the coordinates. Again, it is straightforward to see this using the operators of Lorentz covariant quantization. In this case, the quantum charges are given by (11.76) with $x^\mu(\tau)$ and $p^\mu(\tau)$ taken to be the Heisenberg operators introduced earlier and satisfying the commutation relations (11.58) and (11.59). Both $x^\mu(\tau)$ and $p^\mu(\tau)$ are Hermitian operators. The Lorentz charges $M^{\mu\nu}$ are Hermitian as well:

$$(M^{\mu\nu})^\dagger = p^\nu(\tau)x^\mu(\tau) - p^\mu(\tau)x^\nu(\tau) = M^{\mu\nu}\,, \tag{11.77}$$

since the two constants induced by rearranging the coordinates and momenta back to the original form cancel out.

Quick calculation 11.6 Show that

$$\left[\, M^{\mu\nu}\,,\; x^\rho(\tau)\,\right] = i\,\eta^{\mu\rho}\,x^\nu(\tau) - i\,\eta^{\nu\rho}\,x^\mu(\tau). \tag{11.78}$$

This commutator helps us check that the quantum Lorentz charges generate Lorentz transformations:

$$\begin{aligned}\delta x^\rho(\tau) &= \left[-\frac{i}{2}\,\epsilon_{\mu\nu}M^{\mu\nu}\,,\; x^\rho(\tau)\right] \\ &= \frac{1}{2}\,\epsilon_{\mu\nu}\Big(\eta^{\mu\rho}\,x^\nu(\tau) - \eta^{\nu\rho}\,x^\mu(\tau)\Big) \\ &= \frac{1}{2}\,\epsilon^{\rho\nu}x_\nu(\tau) + \frac{1}{2}\,\epsilon^{\rho\mu}\,x_\mu(\tau) = \epsilon^{\rho\nu}x_\nu(\tau)\,.\end{aligned} \tag{11.79}$$

Equation (11.78) can be used in light-cone *coordinates* by simply using light-cone indices. For example,

$$[M^{-I}, x^+(\tau)\,] = i\eta^{-+}\,x^I(\tau) - i\eta^{I+}x^-(\tau) = -i\,x^I(\tau)\,, \tag{11.80}$$

since $\eta^{I+} = 0$. The operator M^{-I} here is a Lorentz covariant generator expressed in light-cone coordinates. It is *not* a light-cone gauge Lorentz generator. Those we have not yet constructed.

Given a set of quantum operators, it is interesting to calculate their commutators. In quantum mechanics, for example, you learned that the components L_x, L_y, and L_z of the angular momentum satisfy a set of commutation relations ($[L_x, L_y] = iL_z$, and others) that define the Lie algebra of angular momentum. The momentum operators p^μ considered earlier define a very simple Lie algebra: they all commute. We would like to know what is the commutator of two Lorentz generators. The computation takes a few steps (Problem 11.5). Using equation (11.78), and a similar equation for $[M^{\mu\nu}\,,\; p^\rho]$, one finds that the commutator can be written as a linear combination of four Lorentz generators:

$$[M^{\mu\nu}, M^{\rho\sigma}\,] = i\eta^{\mu\rho}M^{\nu\sigma} - i\eta^{\nu\rho}M^{\mu\sigma} + i\eta^{\mu\sigma}M^{\rho\nu} - i\eta^{\nu\sigma}M^{\rho\mu}\,. \tag{11.81}$$

This result defines the Lorentz Lie algebra. Equation (11.81) must be satisfied by the analogous operators $M^{\mu\nu}$ of any Lorentz invariant quantum theory. If it is not possible to construct such operators, the theory is not Lorentz invariant. This will be crucial to our quantization of the string, for requiring that (11.81) holds imposes additional restrictions, which have significant physical consequences.

Quick calculation 11.7 Since $M^{\mu\nu} = -M^{\nu\mu}$, the left-hand side of (11.81) changes sign under the exchange of μ and ν. Verify that the right-hand side also changes sign under this exchange.

We can now use (11.81) to determine the commutators of Lorentz charges in light-cone *coordinates*. The Lorentz generators are given by

$$M^{IJ}, \quad M^{+I}, \quad M^{-I}, \quad \text{and} \quad M^{+-}. \tag{11.82}$$

Consider, for example, the commutator $[M^{+-}, M^{+I}]$. To use (11.81) notice the structure of its right-hand side: each η contains one index from each of the generators in the left-hand side. For $[M^{+-}, M^{+I}]$, the only way to get a nonvanishing η is to use the $-$ from the first generator and the $+$ from the second generator. The nonvanishing term is the second one on the right-hand side of (11.81), and we find

$$[M^{+-}, M^{+I}] = -i\,\eta^{-+} M^{+I} = i M^{+I}. \tag{11.83}$$

Similarly,

$$[M^{-I}, M^{-J}] = 0. \tag{11.84}$$

Here η must use the I and J indices, but then the other two indices must go into M giving us M^{--}, which vanishes by antisymmetry.

Quick calculation 11.8 Show that $M^{+-} = M^{10}$. This shows that M^{+-} generates a boost along the x^1 direction.

So far, we have considered the covariant Lorentz charges in light-cone *coordinates*. We must now find Lorentz charges for our light-cone *gauge* quantization of the particle. Our earlier discussion of the momenta suggests that we really face three questions.

(1) How are these charges going to be defined?
(2) What kind of transformations will they generate?
(3) Which commutation relations will they satisfy?

In the remaining part of this section we will explore question (1) in detail. Before doing so, let us give brief answers to questions (2) and (3), leaving further analysis of these questions to Problems 11.6 and 11.7. The light-cone gauge Lorentz generators are expected to generate Lorentz transformations of coordinates and momentum, but in some cases, these transformations will be accompanied by reparameterizations of the world-line. Regarding (3), the light-cone gauge Lorentz generators will satisfy the same commutation relations that the covariant operators in light-cone coordinates do. This establishes that Lorentz symmetry holds in the light-cone theory of the quantum point particle. The success of the

construction is not obvious *a priori*. It is not clear that the reduced set of light-cone gauge operators suffices to construct quantum Lorentz charges that generate Lorentz transformations (plus other transformations) and satisfy the Lorentz algebra.

The simplest guess for the light-cone gauge generators is to use light-cone coordinates in the covariant formula (11.76) and then replace $x^+(\tau)$, $x^-(\tau)$, and p^- using their light-cone gauge definitions in (11.29), (11.30), and (11.31). Let us try this prescription with M^{+-}:

$$\begin{aligned} M^{+-} &\stackrel{?}{=} x^+(\tau)\, p^-(\tau) - x^-(\tau)\, p^+(\tau) \\ &\stackrel{?}{=} \frac{p^+\tau}{m^2}\, p^- - \Big(x_0^- + \frac{p^-}{m^2}\tau\Big) p^+ \\ &\stackrel{?}{=} -x_0^-\, p^+ \,. \end{aligned} \tag{11.85}$$

Since x_0^- and p^+ are τ-independent, so too is M^{+-}. We have a minor complication, however. The operator M^{+-} is not Hermitian: $(M^{+-})^\dagger - M^{+-} = [x_0^-, p^+] \neq 0$. This failure of Hermiticity illustrates how the use of the light-cone gauge can affect basic properties of operators. The covariant Lorentz generators were automatically Hermitian, the light-cone gauge generators are not. We are therefore motivated to define a Hermitian M^{+-} as

$$M^{+-} = -\frac{1}{2}(x_0^-\, p^+ + p^+ x_0^-) \,. \tag{11.86}$$

We take this to be the light-cone gauge Lorentz generator M^{+-}.

The most complicated of all generators is M^{-I}. It is the most interesting one as well. The prescription used for M^{+-} this time gives

$$\begin{aligned} M^{-I} &\stackrel{?}{=} x^-(\tau)\, p^I - x^I(\tau)\, p^- \\ &\stackrel{?}{=} \Big(x_0^- + \frac{p^-}{m^2}\tau\Big) p^I - \Big(x_0^I + \frac{p^I\tau}{m^2}\Big) p^- \\ &\stackrel{?}{=} x_0^-\, p^I - x_0^I\, p^- . \end{aligned} \tag{11.87}$$

As before, the τ-dependence vanishes, but we are left with a complicated result since p^- is a nontrivial function of the other momenta. We define M^{-I} as the Hermitian version of the operator obtained above:

$$M^{-I} \equiv x_0^-\, p^I - \frac{1}{2}\left(x_0^I\, p^- + p^-\, x_0^I\right) . \tag{11.88}$$

If the light-cone gauge Lorentz charges are to satisfy the Lorentz algebra we must have

$$[\, M^{-I}\,,\; M^{-J}\,] = 0 \,, \tag{11.89}$$

as we noted in (11.84). Does M^{-I}, as defined by (11.88), satisfy this equation? The answer is yes, as you will see for yourself in Problem 11.6. This result is necessary to ensure Lorentz invariance of the quantum theory. All other commutators of Lorentz generators also work out correctly.

The calculation of $[\, M^{-I}\,,\; M^{-J}\,]$ in quantum string theory is fairly complicated, but the answer is very interesting. It turns out that this commutator is zero if and only if the string

propagates in a spacetime of some particular dimension and, furthermore, only if the definition of mass is changed in such a way that we can find massless gauge fields in the spectrum of the open string! String theory is such a constrained theory that it is only Lorentz invariant for a fixed spacetime dimensionality.

Problems

Problem 11.1 Equation of motion for Heisenberg operators.

Assume that the Schrödinger Hamiltonian $H = H(p, q)$ is time independent. In this case the time-independent Schrödinger operator ξ yields a Heisenberg operator $\xi(t) = e^{iHt}\xi e^{-iHt}$. Show that this operator satisfies the equation

$$i\frac{d\xi(t)}{dt} = \Big[\,\xi(t)\,,\; H(p(t), q(t))\Big]\,.$$

This computation proves that equation (11.18) holds for time-independent Hamiltonians.

Problem 11.2 Heisenberg operators and time-dependent Hamiltonians.

When the Schrödinger Hamiltonian $H = H(p, q; t)$ is time dependent, time evolution of states is generated by a unitary operator $U(t)$:

$$|\Psi, t\rangle = U(t)|\Psi\rangle\,, \tag{1}$$

where $U(t)$ bears some nontrivial relation to H. Here $|\Psi\rangle$ denotes the state at zero time and $U(0) = 1$.

(a) Use the Schrödinger equation to show that

$$i\frac{dU(t)}{dt} = HU(t)\,. \tag{2}$$

Let $U \equiv U(t)$, for brevity. Since U^{-1} acting on $|\Psi, t\rangle$ gives a time-independent state, considerations similar to those given for (11.24) lead us to define the Heisenberg operator corresponding to the Schrödinger operator α as

$$\alpha(t) = U^{-1}\alpha\, U\,. \tag{3}$$

(b) Let ξ be a time-independent Schrödinger operator, and let $\xi(t)$ be the corresponding Heisenberg operator, defined using (3). Show that

$$i\frac{d\xi(t)}{dt} = \Big[\,\xi(t)\,,\; H(p(t), q(t); t)\Big]\,.$$

This computation proves that equation (11.18) holds for time-dependent Hamiltonians.

(c) If $[\alpha_1, \alpha_2] = \alpha_3$ holds for Schrödinger operators α_1, α_2, and α_3, show that $[\alpha_1(t), \alpha_2(t)] = \alpha_3(t)$ holds for the corresponding Heisenberg operators.

Problem 11.3 Classical dynamics in Hamiltonian language.

Consider a classical phase space (q, p), a trajectory $(q(t), p(t))$, and an observable $v\,(q(t),\, p(t);\; t)$. From the standard rules of differentiation,

$$\frac{dv}{dt} = \frac{\partial v}{\partial t} + \frac{\partial v}{\partial p}\frac{dp}{dt} + \frac{\partial v}{\partial q}\frac{dq}{dt}\,. \tag{1}$$

With the Poisson bracket defined as

$$\{A,\, B\} = \frac{\partial A}{\partial q}\frac{\partial B}{\partial p} - \frac{\partial A}{\partial p}\frac{\partial B}{\partial q}\,, \tag{2}$$

show that

$$\frac{dv}{dt} = \frac{\partial v}{\partial t} + \{v,\, H\}\,. \tag{3}$$

Comparing this result to (11.19) we see the parallel between the time evolution of a general operator $\mathcal{O}$ and the classical Hamiltonian evolution of an observable v in phase space.

To derive (3) you need the classical equations of motion in Hamiltonian language. These can be obtained by demanding that

$$\int dt\Big(p(t)\dot{q}(t) - H(p(t), q(t); t)\Big)$$

be stationary for independent variations $\delta q(t)$ and $\delta p(t)$.

Problem 11.4 Reparameterization symmetries of the point particle.

Show that the variation $\delta x^\mu(\tau) = \lambda(\tau)\partial_\tau x^\mu(\tau)$ induces a variation δL of the point particle Lagrangian that can be written as

$$\delta L(\tau) = \partial_\tau\big(\lambda(\tau)L(\tau)\big)\,.$$

This proves that the reparameterizations δx^μ are symmetries in the sense defined in Problem 8.9. Show, however, that the charges associated with these reparameterization symmetries vanish. When λ is τ-independent, the reparameterization is an infinitesimal constant τ translation. The conserved charge is then the Hamiltonian. Show directly that the Hamiltonian defined canonically from the point particle Lagrangian vanishes.

Problem 11.5 Lorentz generators and Lorentz algebra.

In this problem we consider the Lorentz covariant charges (11.76).

(a) Calculate the commutator $[\,M^{\mu\nu}\,,\; p^\rho\,]$.
(b) Calculate the commutator $[\,M^{\mu\nu}\,,\; M^{\rho\sigma}\,]$ and verify that (11.81) holds.
(c) Consider the Lorentz algebra in light cone *coordinates*. Give

$$[\,M^{\pm I}\,,\; M^{JK}\,], \quad [\,M^{\pm I}\,,\; M^{\mp J}\,], \quad [M^{+-}\,,\; M^{\pm I}\,], \quad \text{and} \quad [\,M^{\pm I}\,,\; M^{\pm J}\,].$$

Problem 11.6 Light-cone gauge commutator $[M^{-I}, M^{-J}]$ for the particle.

The purpose of the present calculation is to show that

$$[\, M^{-I}\,,\; M^{-J}\,] = 0\,. \tag{1}$$

(a) Verify that the light-cone gauge operator M^{-I} takes the form

$$M^{-I} = (x_0^- p^I - x_0^I p^-) + \frac{i}{2}\,\frac{p^I}{p^+}\,. \tag{2}$$

Set up now the computation of (1) distinguishing the two kinds of terms in (2). Calculate the contributions to the commutator from mixed terms and from the last term.

(b) Complete the computation of (1) by finding the contribution from the first term on the right-hand side of (2).

Problem 11.7 Transformations generated by the light-cone gauge Lorentz generators M^{+-} and M^{-I}.

(a) Calculate the commutator of M^{+-} (defined in (11.86)) with the light-cone coordinates $x^+(\tau)$, $x^-(\tau)$, and $x^I(\tau)$. Show that M^{+-} generates the expected Lorentz transformations of these coordinates.

(b) Calculate the commutator of M^{-I} with the light-cone coordinates $x^+(\tau)$, $x^-(\tau)$, and $x^J(\tau)$. Show that M^{-I} generates the expected Lorentz transformations together with a compensating reparameterization of the world-line. Calculate the parameter λ for this reparameterization. [*Hint:* the reparameterization takes the "hermiticized" form $\delta x^\mu(\tau) = \frac{1}{2}(\lambda\, \partial_\tau x^\mu + \partial_\tau x^\mu\, \lambda)$.]

12 Relativistic quantum open strings

We finally quantize the relativistic open string. We use the light-cone gauge to set up commutation relations and to define a Hamiltonian in the Heisenberg picture. We discover an infinite set of creation and annihilation operators, labeled by an integer and a transverse vector index. The oscillators corresponding to the X^- direction are transverse Virasoro operators. The ambiguities we encounter in defining the quantum theory are fixed by requiring that the theory be Lorentz invariant. Among these ambiguities, the dimensionality of spacetime is fixed to the value 26, and the mass formula is shifted slightly from its classical counterpart such that the spectrum admits massless photon states. The spectrum also contains a tachyon state, which indicates the instability of the D25-brane.

12.1 Light-cone Hamiltonian and commutators

We are at long last in a position to quantize the relativistic string. We have acquired considerable intuition for the dynamics of classical relativistic strings, and we have examined in detail how to quantize the simpler, but still nontrivial, relativistic point particle. Moreover, having taken a brief look into the basics of scalar, electromagnetic, and gravitational quantum fields in the light-cone gauge, we will be able to appreciate the implications of quantum open string theory. In this chapter we will deal with open strings. We will assume throughout the presence of a space-filling D-brane. In the next chapter we will quantize the closed string.

Just as before, we will interpret the classical equations of motion in the light-cone gauge as equations for the appropriate Heisenberg operators. It is therefore necessary for us to review the results of our light-cone analysis of the classical relativistic string.

We found a class of world-sheet parameterizations (9.27) for which the equations of motion are wave equations $\ddot{X}^\mu - X^{\mu''} = 0$. This remarkable simplification came at the expense of two constraints: $(\dot{X} \pm X')^2 = 0$. With these constraints, the momentum densities are simple derivatives of the coordinates:

$$\mathcal{P}^{\sigma\mu} = -\frac{1}{2\pi\alpha'} X^{\mu\prime}, \quad \mathcal{P}^{\tau\mu} = \frac{1}{2\pi\alpha'} \dot{X}^\mu. \tag{12.1}$$

These equations hold in all gauges within the class we considered. In particular, they are true in the light-cone gauge. For open strings in the light-cone gauge, we set $X^+ = 2\alpha' p^+ \tau$

and solved for X^- in terms of the transverse coordinates X^I. Indeed, using (9.65) with $\beta = 2$, we have

$$\dot{X}^- = \frac{1}{2\alpha'}\frac{1}{2p^+}(\dot{X}^I\dot{X}^I + X^{I\prime}X^{I\prime}). \tag{12.2}$$

This gives us an explicit expression for $\mathcal{P}^{\tau-}$:

$$\begin{aligned}\mathcal{P}^{\tau-} &= \frac{1}{2\pi\alpha'}\dot{X}^- = \frac{1}{2\pi\alpha'}\frac{1}{2\alpha'}\frac{1}{2p^+}(2\pi\alpha')^2\left(\mathcal{P}^{\tau I}\mathcal{P}^{\tau I} + \frac{X^{I\prime}X^{I\prime}}{(2\pi\alpha')^2}\right)\\ &= \frac{\pi}{2p^+}\left(\mathcal{P}^{\tau I}\mathcal{P}^{\tau I} + \frac{X^{I\prime}X^{I\prime}}{(2\pi\alpha')^2}\right).\end{aligned} \tag{12.3}$$

These equations will soon become useful.

As a first step in defining a quantum theory of the light-cone relativistic string, we must give the list of Schrödinger operators. Motivated by the list (11.25) of Schrödinger operators for the quantum point particle, we choose our τ-independent Schrödinger operators to be

$$\text{Schrödinger operators: } \left(X^I(\sigma),\quad x_0^-,\quad \mathcal{P}^{\tau I}(\sigma),\quad p^+\right). \tag{12.4}$$

The associated Heisenberg operators are then

$$\text{Heisenberg operators: } \left(X^I(\tau,\sigma),\quad x_0^-(\tau),\quad \mathcal{P}^{\tau I}(\tau,\sigma),\quad p^+(\tau)\right). \tag{12.5}$$

Because the operators (12.4) have no explicit τ dependence, neither do the Heisenberg operators (12.5). As in the case of the point particle, we expect x_0^- and p^+ to be fully τ-independent Heisenberg operators.

Now we set up the commutation relations. For the Schrödinger operators $X^I(\sigma)$ and $\mathcal{P}^{\tau I}(\sigma)$ we must face the fact that these operators have σ dependence. It is reasonable to demand that such operators fail to commute only if they are at the same point along the string. We do not expect (simultaneous) measurements at different points on the string to interfere with each other. Therefore we set

$$\left[\, X^I(\sigma)\,,\mathcal{P}^{\tau J}(\sigma')\,\right] = i\eta^{IJ}\,\delta(\sigma-\sigma')\,. \tag{12.6}$$

Here the delta function is being used to implement the constraint that the commutator must vanish when $\sigma \neq \sigma'$. We had to use a Dirac delta function, as opposed to a Kronecker delta, since σ is a continuous variable. Equation (12.6) is naturally supplemented with the commutation relations

$$\left[\, X^I(\sigma)\,,X^J(\sigma')\,\right] = \left[\,\mathcal{P}^{\tau I}(\sigma)\,,\mathcal{P}^{\tau J}(\sigma')\,\right] = 0\,, \tag{12.7}$$

and

$$\left[x_0^-\,,\ p^+\right] = -i\,. \tag{12.8}$$

The operators x_0^- and p^+ commute with all of the other Schrödinger operators:

$$\left[x_0^-\,, X^I(\sigma)\right] = \left[x_0^-\,, \mathcal{P}^{\tau I}(\sigma)\right] = \left[p^+\,, X^I(\sigma)\right] = \left[p^+\,, \mathcal{P}^{\tau I}(\sigma)\right] = 0\,. \tag{12.9}$$

For the associated Heisenberg operators, the only nonvanishing equal-time commutation relations are therefore

$$\left[X^I(\tau,\sigma)\,, \mathcal{P}^{\tau J}(\tau,\sigma')\right] = i\,\eta^{IJ}\,\delta(\sigma-\sigma')\,, \tag{12.10}$$

as well as

$$\left[x_0^-(\tau)\,,\ p^+(\tau)\right] = -i\,. \tag{12.11}$$

All other commutators vanish:

$$\begin{aligned}\left[X^I(\tau,\sigma), X^J(\tau,\sigma')\right] &= \left[\mathcal{P}^{\tau I}(\tau,\sigma), \mathcal{P}^{\tau J}(\tau,\sigma')\right] = 0\,,\\ \left[x_0^-(\tau)\,, X^I(\tau,\sigma)\right] &= \left[x_0^-(\tau)\,, \mathcal{P}^{\tau I}(\tau,\sigma)\right] = 0\,,\\ \left[p^+(\tau)\,, X^I(\tau,\sigma)\right] &= \left[p^+(\tau)\,, \mathcal{P}^{\tau I}(\tau,\sigma)\right] = 0\,.\end{aligned} \tag{12.12}$$

We must now invent the Hamiltonian. Our Hamiltonian should generate τ translation. From our experience with the point particle, we know that p^- generates X^+ translation. But in the light-cone gauge $X^+ = 2\alpha' p^+\tau$, so

$$\frac{\partial}{\partial\tau} = \frac{\partial X^+}{\partial\tau}\frac{\partial}{\partial X^+} = 2\alpha' p^+\frac{\partial}{\partial X^+}\,. \tag{12.13}$$

It follows that the Hamiltonian that generates change in τ is expected to be

$$H = 2\alpha' p^+ p^- = 2\alpha' p^+\int_0^\pi d\sigma\,\mathcal{P}^{\tau -}\,. \tag{12.14}$$

This will indeed turn out to be the correct string Hamiltonian. Using (12.3), the Hamiltonian can be written more explicitly as the Heisenberg operator

$$H(\tau) = \pi\alpha'\int_0^\pi d\sigma\left(\mathcal{P}^{\tau I}(\tau,\sigma)\mathcal{P}^{\tau I}(\tau,\sigma) + \frac{X^{I\prime}(\tau,\sigma)X^{I\prime}(\tau,\sigma)}{(2\pi\alpha')^2}\right). \tag{12.15}$$

H must generate quantum equations of motion that are operator versions of the classical equations of motion. H is very simple when expressed in terms of the transverse

Virasoro modes of Chapter 9. There we saw that $L_0^\perp = 2\alpha' p^+ p^-$ (equation (9.78)), so (12.14) immediately gives

$$H = L_0^\perp \,. \tag{12.16}$$

This expression for the Hamiltonian is perhaps the most memorable one, though, as we will see later on, the true Hamiltonian is slightly different. The operator products $\mathcal{P}\mathcal{P}$ and $X'X'$ in (12.15) are actually ambiguous operators and need careful definition. Additionally, Lorentz invariance will require the subtraction of a calculable constant from H.

Now that we have a plausible candidate for the Hamiltonian, we have to derive the equations of motion. Any Heisenberg operator $\xi(\tau,\sigma)$ which arises from a time-independent Schrödinger operator $\xi(\sigma)$ must satisfy

$$i\,\dot{\xi}(\tau,\sigma) = [\,\xi(\tau,\sigma)\,,\; H(\tau)\,]\,, \tag{12.17}$$

where $H(\tau)$ is given in (12.15). Since $H(\tau)$ is built from Heisenberg operators that have no explicit time dependence, we can substitute $H(\tau)$ for $\xi(\tau,\sigma)$ in (12.17). We conclude that it is completely time independent: $H(\tau) = H$. Furthermore, we can see that $x_0^-(\tau)$ and $p^+(\tau)$ commute with H. They are therefore time independent operators, so we will henceforth denote them by x_0^- and p^+. The commutator (12.11) then becomes

$$\left[\,x_0^-\,,\; p^+\right] = -i\,. \tag{12.18}$$

The Heisenberg equation of motion for $X^I(\tau,\sigma)$ is

$$i\dot{X}^I(\tau,\sigma) = [X^I(\tau,\sigma)\,,\; H(\tau)] = \left[\,X^I(\tau,\sigma)\,, \pi\alpha' \int_0^\pi d\sigma'\, \mathcal{P}^{\tau J}(\tau,\sigma')\mathcal{P}^{\tau J}(\tau,\sigma')\right],$$

where we dropped the second term in H since it commutes with $X^I(\tau,\sigma)$:

$$\left[X^I(\tau,\sigma)\,,\; X^{J\prime}(\tau,\sigma')\right] = \frac{\partial}{\partial\sigma'}\left[X^I(\tau,\sigma)\,,\; X^J(\tau,\sigma')\right] = 0\,. \tag{12.19}$$

We also reinserted the time parameter in $H(\tau)$, choosing a time that gives easily evaluated equal-time commutators. Making use of (12.10) we find

$$i\dot{X}^I(\tau,\sigma) = \pi\alpha' \cdot 2 \cdot \int_0^\pi d\sigma'\, \mathcal{P}^{\tau J}(\tau,\sigma')\, i\,\eta^{IJ}\delta(\sigma-\sigma')\,. \tag{12.20}$$

Performing the integral and cancelling the common factor of i, we find

$$\dot{X}^I(\tau,\sigma) = 2\pi\alpha'\,\mathcal{P}^{\tau I}(\tau,\sigma)\,. \tag{12.21}$$

Happily, this coincides with the classical equation of motion (12.1). The other equations of motion can be checked in a similar fashion. For example, you can calculate $\dot{\mathcal{P}}^{\tau I}$ and use the result to verify that

$$\ddot{X}^I - X^{I\prime\prime} = 0 \tag{12.22}$$

is the quantum equation of motion (Problem 12.1). As we turn classical string theory into a quantum theory, the classical boundary conditions become operator equations. For example, the Neumann boundary conditions

$$\partial_\sigma X^I(\tau,\sigma) = 0\,, \quad \sigma = 0,\pi\,, \tag{12.23}$$

are taken literally as the condition that the operator $\partial_\sigma X^I(\tau,\sigma)$ vanishes at the open string endpoints.

We learned in Chapter 9 that the linear combinations of derivatives $(\dot{X}^I \pm X^{I\prime})$ are particularly simple and useful. We conclude this section by calculating commutators of these derivatives. We begin by using (12.21) to rewrite the commutator (12.10) as

$$\Big[\, X^I(\tau,\sigma)\,,\ \dot{X}^J(\tau,\sigma')\,\Big] = 2\pi\alpha'\, i\,\eta^{IJ}\,\delta(\sigma-\sigma')\,. \tag{12.24}$$

Taking the σ derivative of this equation yields

$$\Big[\, X^{I\prime}(\tau,\sigma)\,,\ \dot{X}^J(\tau,\sigma')\,\Big] = 2\pi\alpha'\, i\,\eta^{IJ}\,\frac{d}{d\sigma}\delta(\sigma-\sigma')\,. \tag{12.25}$$

Differentiating $\big[\, X^I(\tau,\sigma), X^J(\tau,\sigma')\big] = 0$ with respect to σ and σ' and recalling that $\big[\, \mathcal{P}^{\tau I}(\tau,\sigma), \mathcal{P}^{\tau J}(\tau,\sigma')\big] = 0$, we find that τ and σ derivatives of the coordinates separately commute among themselves:

$$\Big[\, X^{I\prime}(\tau,\sigma)\,,\ X^{J\prime}(\tau,\sigma')\Big] = \Big[\, \dot{X}^I(\tau,\sigma)\,,\ \dot{X}^J(\tau,\sigma')\Big] = 0\,. \tag{12.26}$$

Now we examine the commutator

$$\Big[\, (\dot{X}^I + X^{I\prime})(\tau,\sigma)\,,\ (\dot{X}^J + X^{J\prime})(\tau,\sigma')\,\Big]\,, \tag{12.27}$$

which as a consequence of (12.26) equals

$$\Big[\, \dot{X}^I(\tau,\sigma)\,,\ X^{J\prime}(\tau,\sigma')\,\Big] + \Big[\, X^{I\prime}(\tau,\sigma)\,,\ \dot{X}^J(\tau,\sigma')\,\Big]\,. \tag{12.28}$$

The second term is given by (12.25). The first term equals

$$-\Big[X^{J\prime}(\tau,\sigma')\,,\dot{X}^I(\tau,\sigma)\,\Big] = -(2\pi\alpha')i\eta^{JI}\frac{d}{d\sigma'}\delta(\sigma'-\sigma) = 2\pi\alpha'\, i\eta^{IJ}\frac{d}{d\sigma}\delta(\sigma-\sigma')\,.$$

To obtain this result we noted that a σ' derivative can be traded for minus a σ derivative when it acts on a function of $(\sigma-\sigma')$. Moreover, we used $\delta(x) = \delta(-x)$. We now see that both terms in (12.28) are equal, so

$$\Big[\, (\dot{X}^I + X^{I\prime})(\tau,\sigma)\,,\ (\dot{X}^J + X^{J\prime})(\tau,\sigma')\,\Big] = 4\pi\alpha'\, i\eta^{IJ}\,\frac{d}{d\sigma}\delta(\sigma-\sigma')\,. \tag{12.29}$$

In fact, more generally, we have found that

$$\Big[\, (\dot{X}^I \pm X^{I\prime})(\tau,\sigma)\,,\ (\dot{X}^J \pm X^{J\prime})(\tau,\sigma')\,\Big] = \pm 4\pi\alpha'\, i\eta^{IJ}\,\frac{d}{d\sigma}\delta(\sigma-\sigma')\,, \tag{12.30}$$

since only cross terms contribute. Finally,

$$\Big[\,(\dot{X}^I \pm X^{I\prime})(\tau,\sigma)\,,\;(\dot{X}^J \mp X^{J\prime})(\tau,\sigma')\,\Big] = 0\,. \tag{12.31}$$

Equations (12.30) and (12.31) hold for $\sigma, \sigma' \in [0,\pi]$.

12.2 Commutation relations for oscillators

The commutation relations written so far are delicate to handle because they involve field operators and use delta functions. They are an infinite set of relations which hold for continuous values of σ and σ'. It is therefore useful to recast them in discrete form, namely, as a denumerable set of commutation relations. For this purpose we will examine the mode expansions of Section 9.4. These followed from the classical wave equations and the boundary conditions for a space-filling D-brane. Since the wave equations and the boundary conditions continue to hold in the quantum theory, we can use the mode expansions in the quantum theory. The classical modes α_n^I, however, become quantum operators with nontrivial commutation relations.

Recall our solution (9.69) to the wave equation with Neumann boundary conditions:

$$X^I(\tau,\sigma) = x_0^I + \sqrt{2\alpha'}\,\alpha_0^I\tau + i\sqrt{2\alpha'}\sum_{n\neq 0}\frac{1}{n}\alpha_n^I\cos n\sigma\; e^{-in\tau}\,. \tag{12.32}$$

In addition, from (9.74) we have

$$\begin{aligned}(\dot{X}^I + X^{I\prime})(\tau,\sigma) &= \sqrt{2\alpha'}\sum_{n\in\mathbb{Z}}\alpha_n^I\, e^{-in(\tau+\sigma)}\,, \quad \sigma\in[0,\pi]\,,\\ (\dot{X}^I - X^{I\prime})(\tau,\sigma) &= \sqrt{2\alpha'}\sum_{n\in\mathbb{Z}}\alpha_n^I\, e^{-in(\tau-\sigma)}\,, \quad \sigma\in[0,\pi]\,.\end{aligned} \tag{12.33}$$

The above equalities hold for $\sigma\in[0,\pi]$ because the open string coordinates are only defined for $\sigma\in[0,\pi]$. We will now construct a function of σ with period 2π which is naturally expressed in terms of open string coordinates. To do this, we *evaluate* the second equation above at $-\sigma$:

$$(\dot{X}^I - X^{I\prime})(\tau,-\sigma) = \sqrt{2\alpha'}\sum_{n\in\mathbb{Z}}\alpha_n^I\, e^{-in(\tau+\sigma)}, \quad \sigma\in[-\pi,0]. \tag{12.34}$$

The range $\sigma\in[-\pi,0]$ is required because otherwise the left-hand side is not defined. We now define the operator $A^I(\tau,\sigma)$ by

$$A^I(\tau,\sigma) \equiv \sqrt{2\alpha'}\sum_{n\in\mathbb{Z}}\alpha_n^I\, e^{-in(\tau+\sigma)}\,, \quad A^I(\tau,\sigma+2\pi) = A^I(\tau,\sigma)\,. \tag{12.35}$$

The indicated periodicity is a direct consequence of the definition. It is now simple to relate A^I to open string coordinates over the length-2π interval $\sigma \in [-\pi, \pi]$. For $\sigma \in [0, \pi]$ we use the top equation in (12.33) and for $\sigma \in [-\pi, 0]$ we use (12.34):

$$A^I(\tau,\sigma) = \begin{cases} (\dot{X}^I + X^{I\prime})(\tau,\,\sigma) & \sigma \in [\,0\,,\,\pi\,] \\ (\dot{X}^I - X^{I\prime})(\tau,-\sigma) & \sigma \in [-\pi, 0]\,. \end{cases} \tag{12.36}$$

The operator A^I will be useful to determine the commutation relations for the α_n^I oscillators. For this, we must compute the commutator $[A^I(\tau,\sigma),\, A^J(\tau,\sigma')]$. Given (12.36) the evaluation for the full range $\sigma, \sigma' \in [-\pi, \pi]$ requires four computations:

$$\begin{aligned}
&\big[\,(\dot{X}^I + X^{I\prime})(\tau,\sigma)\,,\ (\dot{X}^J + X^{J\prime})\,(\tau,\,\sigma')\,\big], \quad \sigma,\sigma' \in [0,\pi],\\
&\big[\,(\dot{X}^I + X^{I\prime})(\tau,\sigma)\,,\ (\dot{X}^J - X^{J\prime})(\tau,-\sigma')\,\big], \quad \sigma \in [0,\pi],\ \sigma' \in [-\pi,0],\\
&\big[\,(\dot{X}^I - X^{I\prime})(\tau,-\sigma)\,,\ (\dot{X}^J + X^{J\prime})(\tau,\sigma')\,\big], \quad \sigma \in [-\pi,0],\ \sigma' \in [0,\pi],\\
&\big[\,(\dot{X}^I - X^{I\prime})(\tau,-\sigma)\,,\ (\dot{X}^J - X^{J\prime})(\tau,-\sigma')\,\big], \quad \sigma,\sigma' \in [-\pi,0]\,.
\end{aligned} \tag{12.37}$$

The first commutator, for $\sigma, \sigma' \in [0, \pi]$, is simply read from (12.30):

$$\big[\,(\dot{X}^I + X^{I\prime})(\tau,\sigma)\,,\ (\dot{X}^J + X^{J\prime})\,(\tau,\,\sigma')\,\big] = 4\pi\alpha' i\eta^{IJ}\frac{d}{d\sigma}\delta(\sigma-\sigma'). \tag{12.38}$$

The last commutator, for $\sigma, \sigma' \in [-\pi, 0]$, is also obtained from (12.30):

$$\begin{aligned}
\big[(\dot{X}^I - X^{I\prime})(\tau,-\sigma)\,,\ (\dot{X}^J - X^{J\prime})(\tau,-\sigma')\,\big] &= -4\pi\alpha' i\eta^{IJ}\frac{d}{d(-\sigma)}\delta(-\sigma+\sigma')\\
&= 4\pi\alpha' i\eta^{IJ}\frac{d}{d\sigma}\delta(\sigma-\sigma'),
\end{aligned}$$

and coincides with the result (12.38) of the first commutator. For the second and third commutators in (12.2) we can use (12.31) to conclude that both of them vanish. In fact, the right-hand side of (12.38) also vanishes since in those cases σ and σ' cannot be equal. It follows that the result of all four commutators can be summarized as

$$\big[A^I(\tau,\sigma),\, A^J(\tau,\sigma')\big] = 4\pi\alpha' i\eta^{IJ}\frac{d}{d\sigma}\delta(\sigma-\sigma')\,, \quad \sigma,\sigma' \in [-\pi,\pi]. \tag{12.39}$$

Using (12.35) and cancelling a common factor of $2\alpha'$, the above result gives

$$\sum_{m',n'\in\mathbb{Z}} e^{-im'(\tau+\sigma)}e^{-in'(\tau+\sigma')}\left[\alpha^I_{m'}\,,\alpha^J_{n'}\right] = 2\pi\, i\eta^{IJ}\frac{d}{d\sigma}\delta(\sigma-\sigma'). \tag{12.40}$$

This equation holds for $\sigma, \sigma' \in [-\pi, \pi]$. In order to extract information we apply on both sides the integral operations

$$\frac{1}{2\pi}\int_0^{2\pi} d\sigma\, e^{im\sigma}\cdot\ \frac{1}{2\pi}\int_0^{2\pi} d\sigma'\, e^{in\sigma'}. \tag{12.41}$$

On the left-hand side of (12.40) the integrals pick the term with $m' = m$ and $n' = n$:

$$e^{-i(m+n)\tau}\left[\alpha^I_m\,,\alpha^J_n\right]. \tag{12.42}$$

On the right-hand side of (12.40) the integrals give

$$\begin{aligned} & i\eta^{IJ}\frac{1}{2\pi}\int_0^{2\pi} d\sigma\, e^{im\sigma}\,\frac{d}{d\sigma}\int_0^{2\pi} d\sigma'\, e^{in\sigma'}\delta(\sigma-\sigma') \\ &= i\eta^{IJ}\frac{1}{2\pi}\int_0^{2\pi} d\sigma\, e^{im\sigma}\,\frac{d}{d\sigma}e^{in\sigma} = -n\,\eta^{IJ}\frac{1}{2\pi}\int_0^{2\pi} d\sigma\, e^{i(m+n)\sigma} \\ &= -n\,\eta^{IJ}\delta_{m+n,0} = m\,\eta^{IJ}\delta_{m+n,0}. \end{aligned} \tag{12.43}$$

Equating our results (12.42) and (12.43), we find

$$[\alpha_m^I\,,\alpha_n^J\,] = m\,\eta^{IJ}\delta_{m+n,0}\,e^{+i(m+n)\tau} = m\,\eta^{IJ}\delta_{m+n,0}, \tag{12.44}$$

since the Kronecker delta can be used to set $m=-n$. Therefore, the commutation relation is

$$\left[\,\alpha_m^I\,,\alpha_n^J\,\right] = m\,\eta^{IJ}\delta_{m+n,0}. \tag{12.45}$$

This is the fundamental commutation relation between α modes. Note that α_0^I commutes with all other oscillators. This is quite reasonable: as shown in (9.52), α_0^I is proportional to the momentum of the string

$$\alpha_0^I = \sqrt{2\alpha'}\,p^I\,, \tag{12.46}$$

and it is expected to have a nontrivial commutator with x_0^J only.

To complete the list of all possible commutators we must find the commutators between x_0^I and the oscillators α_n^J. For this, we consider equation (12.24), and integrate both sides of the equation over $\sigma\in[0,\pi]$. On the left-hand side, the terms with oscillators in $X^I(\tau,\sigma)$ give no contribution, and on the right-hand side the delta function disappears giving a factor of one:

$$\left[\,x_0^I+\sqrt{2\alpha'}\,\alpha_0^I\,\tau\,,\ \dot{X}^J(\tau,\sigma')\right] = 2\alpha'\,i\,\eta^{IJ}\ . \tag{12.47}$$

Since $\dot{X}^J$ is a sum of terms that contain α_n^J, we have $[\alpha_0^I\,,\dot{X}^I]=0$. Additionally, using the mode expansion of $\dot{X}^J$, equation (12.47) becomes

$$\sum_{n'\in\mathbb{Z}}\left[\,x_0^I\,,\alpha_{n'}^J\right]\cos n'\sigma'\,e^{-in'\tau} = \sqrt{2\alpha'}\,i\,\eta^{IJ}\,. \tag{12.48}$$

Reorganizing the left-hand side of this equation we find

$$[\,x_0^I\,,\alpha_0^J\,]+\sum_{n'=1}^{\infty}\left[\,x_0^I\ ,\ \alpha_{n'}^J e^{-in'\tau}+\alpha_{-n'}^J e^{in'\tau}\right]\cos n'\sigma = \sqrt{2\alpha'}\,i\,\eta^{IJ}\,. \tag{12.49}$$

We apply to both sides of this equation the integral operation $\frac{1}{\pi}\int_0^\pi d\sigma \cos n\sigma$, with $n \geq 1$. We then find that

$$\left[x_0^I \, , \, \alpha_n^J e^{-in\tau} + \alpha_{-n}^J e^{in\tau} \right] = 0 \, , \tag{12.50}$$

or equivalently

$$\left[x_0^I \, , \alpha_n^J \right] e^{-in\tau} + \left[x_0^I \, , \alpha_{-n}^J \right] e^{in\tau} = 0 \, . \tag{12.51}$$

Since the left-hand side must vanish for all values of τ, each term must vanish separately (prove this!). It follows that

$$\left[x_0^I \, , \alpha_n^J \right] = 0 \quad \text{for} \quad n \neq 0 \, . \tag{12.52}$$

Additionally, equation (12.49) gives

$$\left[x_0^I, \alpha_0^J \right] = \sqrt{2\alpha'}\, i\eta^{IJ} \, . \tag{12.53}$$

This, together with (12.46), gives the expected commutator

$$\left[x_0^I \, , \, p^J \right] = i\eta^{IJ} \, . \tag{12.54}$$

As in familiar quantum mechanics, the operators x_0^I and p^I are Hermitian:

$$(x_0^I)^\dagger = x_0^I \, , \quad (p^I)^\dagger = p^I \, . \tag{12.55}$$

The calculations we performed to obtain the commutation relations took quite a few steps, which we explained in detail. When we discuss closed strings, or open strings on general D-brane configurations, similar computations will be required. We will be able to carry them out using, with minimal modifications, the calculations we just did.

It is useful at this point to examine in detail the commutation relations (12.45) for the α_n^I modes. As we will show below, they are equivalent to those of an infinite set of creation and annihilation operators. To see this, we begin by defining *oscillators*, taking our inspiration from the classical variables introduced in (9.53):

$$\alpha_n^\mu = a_n^\mu \sqrt{n} \, , \quad \alpha_{-n}^\mu = a_n^{\mu *} \sqrt{n} \, , \quad n \geq 1 \, . \tag{12.56}$$

In these equations, both the α and the a are classical variables. Now they become operators. Classical variables that are complex conjugates of each other become operators that are Hermitian conjugates of each other in the quantum theory. We can therefore preserve the first of the above definitions, but the second must be changed. For our light-cone modes $\mu = I$ we take

$$\alpha_n^I = a_n^I \sqrt{n} \quad \text{and} \quad \alpha_{-n}^I = a_n^{I\dagger} \sqrt{n} \, , \quad n \geq 1 \, . \tag{12.57}$$

Note that, with this definition,

$$(\alpha_n^I)^\dagger = \alpha_{-n}^I\,, \quad n \in \mathbb{Z}\,. \tag{12.58}$$

This equation holds for $n = 0$ because α_0^I, being proportional to p^I, is also Hermitian. It is useful to emphasize that, while the α_n^I modes are defined for all integers n, the a_n^I and $a_n^{I\dagger}$ operators are only defined for positive n.

An important consequence of the above Hermiticity properties is that $X^I(\tau,\sigma)$, which used to be real in the classical theory, is now a Hermitian operator.

Quick calculation 12.1 Use the expansion (12.32) and the Hermiticity conditions (12.55) and (12.58) to show that

$$(X^I(\tau,\sigma))^\dagger = X^I(\tau,\sigma)\,. \tag{12.59}$$

The i factor in front of the sum in (12.32) is needed for this calculation to work out.

We can now rephrase the commutation relation for the α modes in terms of the oscillators $(a_n^I\,,\ a_n^{I\dagger})$. For this purpose, rewrite (12.45) as

$$\left[\,\alpha_m^I\,,\alpha_{-n}^J\,\right] = m\,\delta_{m,n}\eta^{IJ}\,. \tag{12.60}$$

When m and n are integers of opposite signs the right-hand side vanishes, and the two operators in the commutator have mode numbers of the same sign. Therefore, we learn that

$$\left[\,a_m^I\,,a_n^J\,\right] = \left[\,a_m^{I\dagger}\,,\ a_n^{J\dagger}\right] = 0\,. \tag{12.61}$$

If both m and n are positive in (12.60) we find

$$\left[\,\sqrt{m}\,a_m^I\,,\sqrt{n}\,a_n^{J\dagger}\,\right] = m\delta_{m,n}\eta^{IJ}\,. \tag{12.62}$$

Moving the square roots to the right-hand side

$$\left[\,a_m^I\,,a_n^{J\dagger}\,\right] = \frac{m}{\sqrt{mn}}\,\delta_{m,n}\eta^{IJ}\,. \tag{12.63}$$

Since the right-hand side vanishes unless $m = n$, it simplifies to

$$\left[\,a_m^I\,,\ a_n^{J\dagger}\,\right] = \delta_{m,n}\,\eta^{IJ}\,. \tag{12.64}$$

This, together with (12.61), shows that $(a_m^I, a_m^{I\dagger})$ satisfy the commutation relations of the canonical annihilation and creation operators of a quantum simple harmonic oscillator. There is a pair of creation and annihilation operators for each value $m \geq 1$ of the mode number and for each transverse light-cone direction I. The commutation relations are diagonal: oscillators corresponding to different mode numbers, or to different light-cone coordinates, commute. If the mode numbers and the coordinate labels agree, the commutator is equal to one. In terms of the α operators, with $n \geq 1$:

$$\begin{aligned} &\alpha_n^I \quad \text{are annihilation operators,} \\ &\alpha_{-n}^I \quad \text{are creation operators } (n \geq 1). \end{aligned} \tag{12.65}$$

For future reference let us rewrite the expansion of $X^I(\tau, \sigma)$ in (12.32) in terms of creation and annihilation operators. Separating out the sum over all integers into sums over positive and negative integers, and using (12.46), we find

$$X^I(\tau, \sigma) = x_0^I + 2\alpha' \, p^I \tau + i\sqrt{2\alpha'} \sum_{n=1}^{\infty} \left(\alpha_n^I \, e^{-in\tau} - \alpha_{-n}^I \, e^{in\tau} \right) \frac{\cos n\sigma}{n} \, . \tag{12.66}$$

Replacing α modes by the corresponding oscillators we obtain

$$X^I(\tau, \sigma) = x_0^I + 2\alpha' \, p^I \tau + i\sqrt{2\alpha'} \sum_{n=1}^{\infty} \left(a_n^I \, e^{-in\tau} - a_n^{I\dagger} \, e^{in\tau} \right) \frac{\cos n\sigma}{\sqrt{n}} \, . \tag{12.67}$$

This is the expansion of the coordinate operator in terms of creation and annihilation operators.

Let us take stock of what we have learned. The list of operators we started with was given in (12.5). We have seen that the operators $X^I(\tau, \sigma)$ and $\mathcal{P}^{\tau I}(\tau, \sigma)$ can be traded for an infinite collection of oscillators, plus pairs of zero modes (x_0^I, p^I). Since the other two operators in the list, x_0^- and p^+, are also zero modes, the full set of basic operators of string theory is a collection of zero modes plus an infinite set of creation and annihilation operators. This result is so important that we will now derive it in a different way, showing explicitly how the quantum simple harmonic oscillators arise.

12.3 Strings as harmonic oscillators

Our aim here is to give a more physical derivation of the results obtained in the previous section. In particular, we will rederive the mode expansion (12.67) and the commutation relations between the operators in that expansion. These results followed from the fundamental commutation relation (12.10) together with the operator equations of motion (12.22) and the operator boundary conditions (12.23). Of these, the commutation relations (12.10) are perhaps the least intuitive, as they involve a delta function. In the derivation below there will be no delta function.

Here is our strategy. We will invent a simple Lagrangian that describes the dynamics of the light-cone coordinates X^I. This is not such a difficult task since we know the equations of motion of the X^I, their boundary conditions, and the definition of the canonical momenta $\mathcal{P}^{\tau I}$. Then we will expand the coordinate $X^I(\tau, \sigma)$ as a function of σ but with τ-dependent expansion coefficients. Using the Lagrangian, we will show that those expansion

coefficients are actually the coordinates of harmonic oscillators that have ever-increasing energy! We will conclude by relating these oscillators to the creation and annihilation operators obtained in the previous analysis.

To set up the notation, we begin by reviewing the basic properties of the quantum harmonic oscillator. Let $q_n(t)$ be the coordinate of a classical simple harmonic oscillator, and let the action be given by

$$S_n = \int L_n(t)\, dt = \int dt \Big(\frac{1}{2n}\, \dot{q}_n^2(t) - \frac{n}{2}\, q_n^2(t) \Big) . \tag{12.68}$$

We recognize this as a harmonic oscillator because the kinetic energy is proportional to the velocity-squared, and the potential energy is proportional to the coordinate-squared. For this Lagrangian, the momentum p_n conjugate to the coordinate q_n is

$$p_n = \frac{\partial L}{\partial \dot{q}_n} = \frac{1}{n}\, \dot{q}_n \, . \tag{12.69}$$

A little calculation now gives the Hamiltonian as

$$H_n(p_n, q_n) = p_n\, \dot{q}_n - L_n = \frac{n}{2}(p_n^2 + q_n^2) \, . \tag{12.70}$$

In this equation, n plays the role of the frequency ω of the harmonic oscillator. To define the quantum oscillator, we introduce Schrödinger operators q_n and p_n, with the canonical commutation relation

$$[\, q_n \, , \, p_n] = i \, . \tag{12.71}$$

Creation and annihilation operators can be introduced as

$$a_n = \frac{1}{\sqrt{2}}(p_n - i q_n) \, , \quad a_n^\dagger = \frac{1}{\sqrt{2}}(p_n + i q_n) \, . \tag{12.72}$$

You should check that as a consequence of (12.71) the creation and annihilation operators satisfy the commutation relation

$$[a_n \, , \, a_n^\dagger] = 1 \, . \tag{12.73}$$

Inverting the relations in (12.72), we find

$$q_n = \frac{i}{\sqrt{2}}(a_n - a_n^\dagger) \, , \quad p_n = \frac{1}{\sqrt{2}}(a_n + a_n^\dagger) \, . \tag{12.74}$$

These can be used to rewrite the Hamiltonian H_n in terms of the creation and annihilation operators. We find the familiar result

$$H_n = n \Big(a_n^\dagger a_n + \frac{1}{2} \Big) \, . \tag{12.75}$$

We can now consider the Heisenberg operators $(a_n(t), a_n^\dagger(t))$ that are associated with the Schrödinger operators $(a_n, a_n^\dagger)$. As emphasized in Section 11.2, the Heisenberg operators satisfy the same commutation relations as the Schrödinger operators:

$$\Big[a_n(t) \, , a_n^\dagger(t) \Big] = 1 \, . \tag{12.76}$$

The Heisenberg equation of motion for $a_n(t)$ is

$$\dot{a}_n(t) = i\ [\, H_n(t)\,,\, a_n(t)\,] = in\left[\, a_n^\dagger(t) a_n(t)\,,\, a_n(t)\right] = -in\, a_n(t)\,. \tag{12.77}$$

This differential equation is solved by

$$a_n(t) = e^{-int} a_n(0) = e^{-int} a_n\,, \tag{12.78}$$

where a_n is the constant Heisenberg operator that equals $a_n(t)$ at $t = 0$. A similar calculation gives

$$a_n^\dagger(t) = e^{int} a_n^\dagger(0) = e^{int} a_n^\dagger\,. \tag{12.79}$$

As you can see, the angular frequency of oscillation is indeed equal to n. Finally, with these results and (12.74), we can find the explicit time dependence of the operator $q_n(t)$:

$$q_n(t) = \frac{i}{\sqrt{2}}\,(a_n(t) - a_n^\dagger(t)\,) = \frac{i}{\sqrt{2}}\left(a_n\, e^{-int} - a_n^\dagger\, e^{int}\right). \tag{12.80}$$

This concludes our review of the quantum simple harmonic oscillator.

We now turn to the discussion of an action that encodes the dynamics of the transverse light-cone coordinates $X^I(\tau, \sigma)$. We claim that the action is simply given by

$$S = \int d\tau d\sigma\, \mathcal{L} = \frac{1}{4\pi\alpha'} \int d\tau \int_0^\pi d\sigma \left(\dot{X}^I\, \dot{X}^I - X^{I'}\, X^{I'}\right). \tag{12.81}$$

This action is much simpler than the Nambu–Goto action; it has no square root, for example. The first term, which contains time derivatives, represents kinetic energy. The second term, which contains spatial derivatives, represents potential energy. The canonical momentum associated with X^I coincides with the momentum density $\mathcal{P}^{\tau I}$:

$$\frac{\partial \mathcal{L}}{\partial \dot{X}^I} = \frac{1}{2\pi\alpha'} \dot{X}^I = \mathcal{P}^{\tau I}\,, \tag{12.82}$$

as we see comparing with (12.1). This confirms that $\mathcal{L}$ is correctly normalized. The equations of motion for X^I follow by variation:

$$\delta S = \frac{1}{2\pi\alpha'} \int d\tau \int_0^\pi d\sigma \left(\partial_\tau(\delta X^I)\, \dot{X}^I - \partial_\sigma(\delta X^I)\, X^{I'}\right). \tag{12.83}$$

Restricting ourselves to variations where the initial and final positions are fixed, we can drop the total τ derivatives and find

$$\delta S = -\frac{1}{2\pi\alpha'} \int d\tau \left[\, (X^{I'} \delta X^I)\Big|_0^\pi + \int_0^\pi d\sigma\, \delta X^I \left(\ddot{X}^I - X^{I''}\right)\right]. \tag{12.84}$$

It is clear that the requirement that the action be stationary gives us both the wave equation (12.22) for the coordinates and the boundary conditions at the string endpoints. As a final check of the consistency of the action we calculate the Hamiltonian:

$$H = \int_0^\pi d\sigma\, \mathcal{H} = \int_0^\pi d\sigma \left(\mathcal{P}^{\tau I} \dot{X}^I - \mathcal{L}\right). \tag{12.85}$$

Writing the τ derivative of X^I in terms of $\mathcal{P}^{\tau I}$ we find

$$H = \int_0^\pi d\sigma \Big(\pi\alpha' \mathcal{P}^{\tau I} \mathcal{P}^{\tau I} + \frac{1}{4\pi\alpha'} X^{I\prime} X^{I\prime} \Big) . \tag{12.86}$$

This Hamiltonian coincides with the one we postulated and tested in Section 12.1.

Let us now use the action (12.81) to quantize the theory. For this purpose we replace the dynamical variable $X^I(\tau, \sigma)$ by a collection of dynamical variables that have no σ dependence. This is done by writing the expansion

$$X^I(\tau, \sigma) = q^I(\tau) + 2\sqrt{\alpha'} \sum_{n=1}^{\infty} q_n^I(\tau) \frac{\cos n\sigma}{\sqrt{n}} . \tag{12.87}$$

This is the most general expression that satisfies Neumann boundary conditions at the endpoints. The particular normalization used to introduce the expansion coefficients was chosen for convenience.

Our next step is to evaluate the action (12.81) using the above expansion for $X^I(\tau, \sigma)$. For this we use

$$\begin{aligned} \dot{X}^I &= \dot{q}^I(\tau) + 2\sqrt{\alpha'} \sum_{n=1}^{\infty} \dot{q}_n^I(\tau) \frac{\cos n\sigma}{\sqrt{n}} , \\ X^{I\prime} &= -2\sqrt{\alpha'} \sum_{n=1}^{\infty} q_n^I(\tau) \sqrt{n} \, \sin n\sigma . \end{aligned} \tag{12.88}$$

The evaluation of the action S using the above expansions is quite straightforward because the σ integrals of $(\cos n\sigma \cos m\sigma)$ and $(\sin n\sigma \sin m\sigma)$ vanish unless $n = m$. We find,

$$S = \int d\tau \Big[\frac{1}{4\alpha'} \dot{q}^I(\tau) \dot{q}^I(\tau) + \sum_{n=1}^{\infty} \Big(\frac{1}{2n} \dot{q}_n^I(\tau) \dot{q}_n^I(\tau) - \frac{n}{2} q_n^I(\tau) q_n^I(\tau) \Big) \Big] . \tag{12.89}$$

Quick calculation 12.2 Prove equation (12.89).

Comparing the action (12.89) with the one recorded in (12.68) we see that the $q_n^I(\tau)$, with $n \geq 1$, are the coordinates of simple harmonic oscillators. The frequency of oscillation of $q_n^I(\tau)$ is n. This is the physical interpretation of the expansion coefficients in (12.87). Since the action for $q_n^I(\tau)$ coincides exactly with the action S_n, no new work is necessary to compute the Hamiltonian, except for the zero mode q^I:

$$p^I = \frac{\partial L}{\partial \dot{q}^I} = \frac{1}{2\alpha'} \dot{q}^I \quad \text{and} \quad [\, q^I \, , \, p^J \,] = i\eta^{IJ} . \tag{12.90}$$

The Hamiltonian is then given by

$$H = \alpha' \, p^I \, p^I + \sum_{n=1}^{\infty} \frac{n}{2} \Big(p_n^I \, p_n^I + q_n^I \, q_n^I \Big) , \tag{12.91}$$

where we used (12.70) to write the part of the Hamiltonian that arises from the oscillators. The earlier analysis of the Heisenberg operator $q_n(t)$ led to the solution (12.80). This means that for the $q_n^I(\tau)$ oscillators we have

$$q_n^I(\tau) = \frac{i}{\sqrt{2}}\left(a_n^I\, e^{-in\tau} - a_n^{I\dagger} e^{in\tau}\right), \tag{12.92}$$

where $(a_n^I, a_n^{I\dagger})$ are canonically normalized annihilation and creation operators. For the Heisenberg operator $q^I(\tau)$ we find

$$\dot{q}^I(\tau) = i\,[\,H, q^I(\tau)] = i\alpha'\,[\,p^J\; p^J(\tau)\,, q^I(\tau)] = 2\alpha'\; p^I(\tau)\,. \tag{12.93}$$

Note that p^I is a τ-independent Heisenberg operator. We solve this differential equation for $q^I(\tau)$ by writing

$$q^I(\tau) = x_0^I + 2\alpha'\; p^I\;\tau\,. \tag{12.94}$$

Here x_0^I is a constant operator that on account of (12.90) satisfies $[x_0^I,\, p^J] = i\eta^{IJ}$. Finally, we can substitute our solutions (12.92) and (12.94) into the expansion (12.87) for X^I to find

$$X^I(\tau,\sigma) = x_0^I + 2\alpha'\; p^I\,\tau + i\sqrt{2\alpha'}\sum_{n=1}^{\infty}\left(a_n^I\, e^{-in\tau} - a_n^{I\dagger}\, e^{in\tau}\right)\frac{\cos n\sigma}{\sqrt{n}}\,, \tag{12.95}$$

in exact agreement with the previously derived (12.67). We have therefore given a physical derivation of the mode expansion and commutation relations. We identified the classical variables that become oscillators, and we did not have to use delta functions. Having done so in this case, when we quantize other string configurations we will simply use the abstract approach of the previous section. It gives a direct and quick route to the desired answers.

12.4 Transverse Virasoro operators

We have written mode expansions for the transverse coordinates $X^I(\tau,\sigma)$ and we have seen quite explicitly the connection to harmonic oscillators. How about the other light-cone coordinates, $X^+(\tau,\sigma)$ and $X^-(\tau,\sigma)$? The expansion of X^+ is truly simple:

$$X^+(\tau,\sigma) = 2\alpha' p^+\,\tau = \sqrt{2\alpha'}\,\alpha_0^+\,\tau\,. \tag{12.96}$$

As discussed for the classical case in Section 9.5, this means that we are setting

$$x_0^+ = 0\,, \quad \alpha_n^+ = 0\,, \quad n \neq 0\,. \tag{12.97}$$

For the X^- coordinate a mode expansion was provided by (9.72):

$$X^-(\tau,\sigma) = x_0^- + \sqrt{2\alpha'}\,\alpha_0^-\tau + i\sqrt{2\alpha'}\sum_{n\neq 0}\frac{1}{n}\alpha_n^- e^{-in\tau}\cos n\sigma\,. \tag{12.98}$$

Moreover, we used the constraints to solve for X^- in terms of X^I, p^+, and a constant of integration x_0^-. This meant that the α_n^- modes could be written in terms of the α_n^I modes, as shown in equation (9.77):

$$\sqrt{2\alpha'}\,\alpha_n^- = \frac{1}{p^+} L_n^\perp \,, \tag{12.99}$$

where

$$L_n^\perp \equiv \frac{1}{2}\sum_{p\in\mathbb{Z}} \alpha_{n-p}^I \alpha_p^I \,. \tag{12.100}$$

The repeated index I is summed over the transverse light-cone directions. In Chapter 9 the $L_n^\perp$ were called transverse Virasoro *modes*. Having seen that the α modes become operators, the $L_n^\perp$ will now be called transverse Virasoro *operators*. The steps that led to (12.100) remain valid in the quantum theory, except for the fact that the α modes were treated as commuting classical variables. We now know that the α operators do not commute. We must therefore question whether the ordering of the two α operators appearing in (12.100) is the correct one. A better question is whether the ordering matters. Since two α operators fail to commute only when their mode numbers add up to zero, the two operators in $L_n^\perp$ fail to commute only when $n = 0$. So $L_0^\perp$ is the only ambiguous operator.

There is plenty at stake in ordering $L_0^\perp$ correctly. The operator $L_0^\perp$ is in fact the light-cone Hamiltonian, as we showed in equation (12.16). Moreover, we saw at the end of Chapter 9 that $L_0^\perp$ enters directly into the calculation of the mass of string states. We also mentioned at that time that the quantum theory would bring a subtlety into the calculation of the mass. Well, the subtlety has arrived: we must define the quantum operator $L_0^\perp$! Let us therefore look at $L_0^\perp$ in more detail:

$$L_0^\perp = \frac{1}{2}\sum_{p\in\mathbb{Z}} \alpha_{-p}^I \alpha_p^I = \frac{1}{2}\alpha_0^I \alpha_0^I + \frac{1}{2}\sum_{p=1}^{\infty} \alpha_{-p}^I \alpha_p^I + \frac{1}{2}\sum_{p=1}^{\infty} \alpha_p^I \alpha_{-p}^I \,. \tag{12.101}$$

The first sum on the right-hand side is normal-ordered: annihilation operators appear to the right of the creation operators. It is useful to work with normal-ordered operators since they act in a simple manner on the vacuum state. We cannot use operators that do not have a well defined action on the vacuum state. Since the last sum on the right-hand side of (12.101) is not a normal-ordered operator, we rewrite it as

$$\begin{aligned}
\frac{1}{2}\sum_{p=1}^{\infty} \alpha_p^I \alpha_{-p}^I &= \frac{1}{2}\sum_{p=1}^{\infty}\Big(\alpha_{-p}^I \alpha_p^I + [\alpha_p^I \,, \alpha_{-p}^I]\Big) \\
&= \frac{1}{2}\sum_{p=1}^{\infty} \alpha_{-p}^I \alpha_p^I + \frac{1}{2}\sum_{p=1}^{\infty} p\,\eta^{II} \\
&= \frac{1}{2}\sum_{p=1}^{\infty} \alpha_{-p}^I \alpha_p^I + \frac{1}{2}(D-2)\sum_{p=1}^{\infty} p \,.
\end{aligned} \tag{12.102}$$

If you look at the last term of the above equation you will note that it is divergent; it involves the sum of all positive integers! This is clearly problematic. How do we deal with this? One option is simply to ignore this difficulty, claiming that it is really up to us how we define $L_0^\perp$. There is a kernel of truth to this option, but it is not completely correct. Adding a constant to $L_0^\perp$ changes the values of the masses of the string states, and if anything, the above computation has alerted us to the fact that this additive constant could be nonzero, or even infinite. Taken at face value, the above computation gives

$$L_0^\perp = \frac{1}{2}\alpha_0^I\alpha_0^I + \sum_{p=1}^{\infty}\alpha_{-p}^I\alpha_p^I + \frac{1}{2}(D-2)\sum_{p=1}^{\infty} p\,. \tag{12.103}$$

The operator $L_0^\perp$ enters into our computation of the mass via the definition of p^-. From (12.99), with $n = 0$,

$$\sqrt{2\alpha'}\,\alpha_0^- = 2\alpha' p^- = \frac{1}{p^+}L_0^\perp\,. \tag{12.104}$$

This suggests a strategy. First, we *define*, once and for all, $L_0^\perp$ to be the normal-ordered operator in (12.103) *without* including the ordering constant:

$$L_0^\perp \equiv \frac{1}{2}\alpha_0^I\alpha_0^I + \sum_{p=1}^{\infty}\alpha_{-p}^I\alpha_p^I = \alpha'\,p^I\,p^I + \sum_{p=1}^{\infty} p\,a_p^{I\dagger}\,a_p^I\,. \tag{12.105}$$

Note that $L_0^\perp$ is Hermitian: $(L_0^\perp)^\dagger = L_0^\perp$. Second, we introduce an ordering constant a into (12.104):

$$2\alpha'\,p^- \equiv \frac{1}{p^+}\left(L_0^\perp + a\right)\,. \tag{12.106}$$

If we took seriously our attempt to order $L_0^\perp$, we would have to conclude that

$$a \stackrel{?}{=} \frac{1}{2}(D-2)\sum_{p=1}^{\infty} p\,. \tag{12.107}$$

We will discuss below one remarkable interpretation of this equation which does, in fact, give the correct result. More pragmatically, we will take a to be an undetermined constant. As we will show in Section 12.5, the quantum consistency of string theory fixes the constant a to an interesting finite value. Before proceeding further, let us investigate how the inclusion of a modifies the computation of the mass-squared operator. Working from the definition $M^2 = -p^2$, and using (12.105) and (12.106), we find

$$\begin{aligned} M^2 = -p^2 &= 2p^+p^- - p^I\,p^I = \frac{1}{\alpha'}(L_0^\perp + a\,) - p^I\,p^I \\ &= \frac{1}{\alpha'}\Big(a + \sum_{n=1}^{\infty} n a_n^{I\dagger} a_n^I\Big). \end{aligned} \tag{12.108}$$

As expected, a introduces a constant shift into the mass-squared operator.

It is impossible to resist the temptation to interpret (12.107). An important result in mathematics suggests a finite value for the right-hand side. For this we consider the zeta function $\zeta(s)$, which is defined as the infinite sum

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}\,, \quad \Re(s) > 1\,. \tag{12.109}$$

The argument s of the zeta function is assumed to be a complex number, but as indicated, the above sum only converges if the real part of the argument is greater than one. We can use analytic continuation to define the zeta function for all possible values of the argument. $\zeta(s)$ turns out to be finite for all values of s except $s = 1$. In particular, as you will see in Problem 12.4, $\zeta(-1) = -1/12$. On account of (12.109) this suggests that

$$\zeta(-1) = -\frac{1}{12} \stackrel{?}{=} 1 + 2 + 3 + 4 + \cdots. \tag{12.110}$$

This is a surprising interpretation for the infinite sum $\sum_{p=1}^{\infty} p$. Not only is the result finite, but it is also negative! Substituting back in (12.107), it gives us

$$a = -\frac{1}{24}(D-2)\,. \tag{12.111}$$

This is actually the correct value of a, as we will explain in Section 12.5. The inspired guess gave the right answer. We will also see that consistency requires $D = 26$, so that, in fact, $a = -1$. This value for the shift in the mass-squared operator is precisely what is needed for the spectrum of open strings to include massless photon states!

Having discussed $L_0^{\perp}$ in detail, let us consider the other transverse Virasoro operators. Since $(\alpha_n^I)^\dagger = \alpha_{-n}^I$, we may expect that $(\alpha_n^-)^\dagger = \alpha_{-n}^-$, or equivalently, on account of (12.99), that

$$(L_n^{\perp})^\dagger = L_{-n}^{\perp}\,. \tag{12.112}$$

We have already verified this equation for $n = 0$. For $n \neq 0$, we can easily prove this Hermiticity property using (12.100):

$$(L_n^{\perp})^\dagger = \frac{1}{2}\sum_{p\in\mathbb{Z}} (\alpha_{n-p}^I \alpha_p^I)^\dagger = \frac{1}{2}\sum_{p\in\mathbb{Z}} (\alpha_p^I)^\dagger (\alpha_{n-p}^I)^\dagger = \frac{1}{2}\sum_{p\in\mathbb{Z}} \alpha_{-p}^I \alpha_{-n+p}^I\,. \tag{12.113}$$

Since the oscillators in each term of the sum commute, we can exchange them. By also letting $p \to -p$, we get the expected result:

$$(L_n^{\perp})^\dagger = \frac{1}{2}\sum_{p\in\mathbb{Z}} \alpha_{-n-p}^I \alpha_p^I = L_{-n}^{\perp}\,. \tag{12.114}$$

Perhaps the most interesting property of the Virasoro operators is that they do not commute. We have seen that α_m^I and α_n^I commute except when $m + n$ equals zero. This is not the case with the α_n^- modes. Two Virasoro operators $L_m^{\perp}$ and $L_n^{\perp}$ never commute when $m \neq n$. The commutation properties of Virasoro operators are a bit intricate, so we will consider them in steps of increasing generality.

As a warmup, let us consider the commutator between a Virasoro operator and an oscillator α_n^J. We have

$$\left[L_m^\perp , \alpha_n^J \right] = \frac{1}{2} \sum_{p\in\mathbb{Z}} \left[\alpha_{m-p}^I \alpha_p^I , \alpha_n^J \right] = \frac{1}{2} \sum_{p\in\mathbb{Z}} \Big(\alpha_{m-p}^I \left[\alpha_p^I , \alpha_n^J \right] + \left[\alpha_{m-p}^I , \alpha_n^J \right] \alpha_p^I \Big) . \tag{12.115}$$

Evaluating the commutators and recalling that $\eta^{IJ} = \delta^{IJ}$, we find

$$\left[L_m^\perp , \alpha_n^J \right] = \frac{1}{2} \sum_{p\in\mathbb{Z}} \Big(p\, \delta_{p+n,0}\, \alpha_{m-p}^J + (m-p) \delta_{m-p+n,0}\, \alpha_p^J \Big) . \tag{12.116}$$

Because of the Kronecker deltas only one term contributes in each sum: $p = -n$ in the first and $p = m + n$ in the second. We thus find

$$\left[L_m^\perp , \alpha_n^J \right] = \frac{1}{2} \Big(-n\, \alpha_{m+n}^J - n\, \alpha_{m+n}^J \Big) . \tag{12.117}$$

Our final result is therefore

$$\left[L_m^\perp , \alpha_n^J \right] = -n\, \alpha_{m+n}^J . \tag{12.118}$$

The mode number on the right-hand side is the sum of the mode numbers on the left-hand side. This could not be otherwise, since the basic commutator of α operators trades two operators with opposite mode number for a constant, and in doing so, the total mode number is conserved. Moreover, the spatial index on the oscillator is preserved. Equation (12.118) holds for all values of m, including $m = 0$. Indeed, (12.100), which we used, gives $L_0^\perp$ up to a constant that, although infinite, cannot affect the commutator. It is worth anyway to check the result directly.

Quick calculation 12.3 Calculate $[L_0^\perp, \alpha_n^J]$ using (12.105) and confirm that (12.118) holds for $m = 0$.

Quick calculation 12.4 Show that

$$\left[L_m^\perp , x_0^I \right] = -i\sqrt{2\alpha'}\, \alpha_m^I . \tag{12.119}$$

Let us now consider the commutator of two Virasoro operators $L_m^\perp$ and $L_n^\perp$. The computation is a bit subtle, and it is easy to get incorrect answers. We will avoid subtleties by checking, at every stage of the calculation, that our expressions are normal ordered. For this, we begin by rewriting the Visaroso operator (12.100) in such a way that it gives the correct result even for L_0. For this, the sum is split as:

$$L_m^\perp = \frac{1}{2} \sum_{k\geq 0} \alpha_{m-k}^I \alpha_k^I + \frac{1}{2} \sum_{k<0} \alpha_k^I \alpha_{m-k}^I . \tag{12.120}$$

For any value of m the above right-hand side is normal-ordered: in the first sum the α to the right is an annihilation operator (or a zero mode) and in the second sum the α to the left is a creation operator. We now evaluate

$$\begin{aligned}\big[L_m^\perp, L_n^\perp\big] &= \frac{1}{2}\sum_{k\geq 0}\big[\alpha^I_{m-k}\alpha^I_k\,,\; L_n^\perp\big] + \frac{1}{2}\sum_{k<0}\big[\alpha^I_k\alpha^I_{m-k}\,,\; L_n^\perp\big]\,,\\ &= \frac{1}{2}\sum_{k\geq 0}\big[\alpha^I_{m-k},\; L_n^\perp\big]\alpha^I_k \;+\frac{1}{2}\sum_{k<0}\alpha^I_k\big[\alpha^I_{m-k}\,,\; L_n^\perp\big]\\ &\quad+\frac{1}{2}\sum_{k\geq 0}\alpha^I_{m-k}\big[\alpha^I_k\,,\; L_n^\perp\big] + \frac{1}{2}\sum_{k<0}\big[\alpha^I_k\,,\; L_n^\perp\big]\alpha^I_{m-k}\,. \end{aligned} \tag{12.121}$$

Evaluating the commutators we get

$$\begin{aligned}\big[L_m^\perp, L_n^\perp\big] = \;& \frac{1}{2}\sum_{k\geq 0}(m-k)\alpha^I_{m+n-k}\alpha^I_k + \frac{1}{2}\sum_{k<0}(m-k)\alpha^I_k\alpha^I_{m+n-k}\\ &+\frac{1}{2}\sum_{k\geq 0}k\,\alpha^I_{m-k}\alpha^I_{k+n} + \frac{1}{2}\sum_{k<0}k\,\alpha^I_{k+n}\alpha^I_{m-k}\,. \end{aligned} \tag{12.122}$$

The terms on the first line of the right-hand side are always normal ordered. The terms on the second line can require ordering, depending on the values of m and n. Let us now consider two cases: $m+n\neq 0$ and $m+n=0$.

Case m+n ≠ 0. All pairs of oscillators on the right-hand side of (12.122) commute so we find

$$\begin{aligned}\big[L_m^\perp, L_n^\perp\big] &= \frac{1}{2}\sum_{k\in\mathbb{Z}}(m-k)\alpha^I_{m+n-k}\alpha^I_k + \frac{1}{2}\sum_{k\in\mathbb{Z}}k\,\alpha^I_{m-k}\alpha^I_{k+n}\,,\\ &= \frac{1}{2}\sum_{k\in\mathbb{Z}}(m-k)\alpha^I_{m+n-k}\alpha^I_k + \frac{1}{2}\sum_{k\in\mathbb{Z}}(k-n)\,\alpha^I_{m+n-k}\alpha^I_k\,,\\ &= (m-n)\,\frac{1}{2}\sum_{k\in\mathbb{Z}}\alpha^I_{m+n-k}\alpha^I_k\,. \end{aligned} \tag{12.123}$$

In passing from the first to the second line we let $k\to k-n$ in the second sum. Since $m+n\neq 0$ the final operator requires no ordering and is recognized to be $L^\perp_{m+n}$. We have therefore shown that

$$\Big[\,L_m^\perp, L_n^\perp\,\Big] = (m-n)L^\perp_{m+n}\,,\quad m+n\neq 0\,. \tag{12.124}$$

The commutator of two Virasoro operators is a Virasoro operator with mode number equal to the sum of the mode numbers of the operators which enter the commutator. The above result does not hold when $m+n=0$, in which case the answer is different. Nevertheless, as a mathematical construct, a set of operators $L_n^\perp$ with $n\in\mathbb{Z}$, satisfying (12.124) for *all* m and n, defines an interesting Lie algebra (Problem 12.5). This algebra is called the *Virasoro algebra without central extension* or the Witt algebra.

Case m+n =0. In this case we write $n = -m$ and equation (12.122) becomes

$$\begin{aligned}\left[L_m^\perp, L_{-m}^\perp\right] = &\frac{1}{2}\sum_{k=0}^{\infty}(m-k)\alpha_{-k}^I\alpha_k^I + \frac{1}{2}\sum_{k<0}(m-k)\alpha_k^I\alpha_{-k}^I \\ &+ \frac{1}{2}\sum_{k=0}^{\infty}k\,\alpha_{m-k}^I\alpha_{k-m}^I + \frac{1}{2}\sum_{k<0}k\,\alpha_{k-m}^I\alpha_{m-k}^I\,. \end{aligned} \tag{12.125}$$

In order to compare the various terms it is useful to relabel the summation variables so that the rightmost oscillator is always α_k^I. For this we let $k \to -k$ in the second term of the first line, $k \to m+k$ in the first term of the second line, and $k \to m-k$ in the second term of the second line:

$$\begin{aligned}\left[L_m^\perp, L_{-m}^\perp\right] = &\frac{1}{2}\sum_{k=0}^{\infty}(m-k)\alpha_{-k}^I\alpha_k^I + \frac{1}{2}\sum_{k=1}^{\infty}(m+k)\alpha_{-k}^I\alpha_k^I \\ &+ \underline{\frac{1}{2}\sum_{k=-m}^{\infty}(m+k)\,\alpha_{-k}^I\alpha_k^I} + \frac{1}{2}\sum_{k=m+1}^{\infty}(m-k)\,\alpha_{-k}^I\alpha_k^I\,. \end{aligned} \tag{12.126}$$

We now assume, without loss of generality, that $m > 0$. It then follows that all terms are normal-ordered except those in the underlined summand for which $-m \le k \le 0$. Splitting the sum, the underlined term becomes equal to

$$\begin{aligned}&\frac{1}{2}\sum_{k=0}^{m}(m-k)\,\alpha_k^I\alpha_{-k}^I + \frac{1}{2}\sum_{k=1}^{\infty}(m+k)\,\alpha_{-k}^I\alpha_k^I \\ &= \frac{1}{2}\sum_{k=0}^{m}(m-k)\,[\alpha_k^I, \alpha_{-k}^I] + \frac{1}{2}\sum_{k=0}^{m}(m-k)\,\alpha_{-k}^I\alpha_k^I + \frac{1}{2}\sum_{k=1}^{\infty}(m+k)\,\alpha_{-k}^I\alpha_k^I\,. \end{aligned}$$

Evaluating the commutator and substituting the result back into (12.126) we get

$$\left[L_m^\perp, L_{-m}^\perp\right] = \sum_{k=0}^{\infty}(m-k)\alpha_{-k}^I\alpha_k^I + \sum_{k=1}^{\infty}(m+k)\alpha_{-k}^I\alpha_k^I + (D-2)A(m)\,, \tag{12.127}$$

where $A(m)$ is the constant

$$A(m) = \frac{1}{2}\sum_{k=0}^{m}k(m-k) = \frac{1}{2}m\sum_{k=1}^{m}k - \frac{1}{2}\sum_{k=1}^{m}k^2\,. \tag{12.128}$$

To evaluate $A(m)$ we need the following result:

Quick calculation 12.5 Use mathematical induction to prove that

$$\sum_{k=1}^{m}k^2 = \frac{1}{6}(2m^3 + 3m^2 + m)\,. \tag{12.129}$$

Making use of (12.129) we find

$$A(m) = \frac{1}{4}m^2(m+1) - \frac{1}{12}(2m^3 + 3m^2 + m) = \frac{1}{12}(m^3 - m)\,. \tag{12.130}$$

Expanding out the sums in (12.127) and substituting the value of $A(m)$, we find

$$\left[L_m^\perp, L_{-m}^\perp\right] = 2m\Big(\frac{1}{2}\alpha_0^I\alpha_0^I + \sum_{k=0}^{\infty}\alpha_{-k}^I\alpha_k^I\Big) + \frac{1}{12}(D-2)(m^3-m)\,. \tag{12.131}$$

We recognize that the operator in parenthesis is $L_0^\perp$. Therefore, our final result is

$$\left[L_m^\perp, L_{-m}^\perp\right] = 2m\,L_0^\perp + \frac{1}{12}(D-2)(m^3-m)\,. \tag{12.132}$$

This completes the computation for the case $m+n=0$.

The general result for the commutator of two Virasoro operators is obtained by writing a formula that gives (12.124) for $m+n\neq 0$ and gives (12.132) for $m+n=0$. This is easily done and the result is

$$\Big[L_m^\perp, L_n^\perp\Big] = (m-n)\,L_{m+n}^\perp + \frac{D-2}{12}\Big(m^3-m\Big)\,\delta_{m+n,0}\,. \tag{12.133}$$

A set of operators $L_n^\perp$, with $n\in\mathbb{Z}$, satisfying (12.133) defines the *centrally extended Virasoro algebra*. The second term on the above right-hand side is called the central extension. It is a constant or, more properly, a constant times the identity operator (the operator that acting on any state gives back the state). This term is said to be central because it commutes with all other operators in the algebra. The central term vanishes for $m=0$ and for $m=\pm 1$. There is therefore no central term in the commutator $[L_1, L_{-1}]$. The Virasoro algebra is perhaps the most important algebra in string theory. In the light-cone quantization of string theory – our subject in this chapter – the Virasoro operators enter into the definition of the Lorentz generators, as we will see in Section 12.5.

We conclude this section by studying how the Virasoro operators act on the string coordinates. Since quantum operators act via commutators, we must find the commutator of a Virasoro operator with the coordinate operator $X^I(\tau,\sigma)$. We will see that the Virasoro operators generate reparameterizations of the world-sheet.

Making use of the coordinate expansion (12.32) we find

$$\begin{aligned}[\,L_m^\perp\,, X^I(\tau,\sigma)\,] &= [L_m^\perp, x_0^I] + i\sqrt{2\alpha'}\sum_{n\neq 0}\frac{1}{n}\cos n\sigma\, e^{-in\tau}[L_m^\perp, \alpha_n^I] \\ &= -i\sqrt{2\alpha'}\alpha_m^I - i\sqrt{2\alpha'}\sum_{n\neq 0}\cos n\sigma\, e^{-in\tau}\,\alpha_{m+n}^I\,,\end{aligned} \tag{12.134}$$

where we used (12.118) and (12.119) to evaluate the commutators. The right-hand side above can be written as a single sum:

$$\begin{aligned}[L_m^\perp , X^I(\tau,\sigma)] &= -i\sqrt{2\alpha'} \sum_{n\in\mathbb{Z}} \cos n\sigma \, e^{-in\tau} \, \alpha^I_{m+n} \\ &= -i\sqrt{2\alpha'} \, \frac{1}{2} \sum_{n\in\mathbb{Z}} (e^{-in(\tau-\sigma)} + e^{-in(\tau+\sigma)}) \, \alpha^I_{m+n} \\ &= -i\sqrt{2\alpha'} \, \frac{1}{2} \sum_{n\in\mathbb{Z}} (e^{-i(n-m)(\tau-\sigma)} + e^{-i(n-m)(\tau+\sigma)}) \, \alpha^I_{n} \, ,\end{aligned}$$

where in the last step we let $n \to n - m$. Finally,

$$\begin{aligned}[L_m^\perp , X^I(\tau,\sigma)] = &-\frac{i}{2} \, e^{im(\tau-\sigma)} \, \sqrt{2\alpha'} \sum_{n\in\mathbb{Z}} e^{-in(\tau-\sigma)} \alpha^I_n \\ &- \frac{i}{2} \, e^{im(\tau+\sigma)} \sqrt{2\alpha'} \sum_{n\in\mathbb{Z}} e^{-in(\tau+\sigma)} \, \alpha^I_n .\end{aligned}$$

To interpret this result it is necessary to express the right-hand side in terms of derivatives of the string coordinates. This is done with the help of (12.33):

$$\begin{aligned}[L_m^\perp , X^I(\tau,\sigma)] &= -\frac{i}{2} e^{im(\tau-\sigma)} \, (\dot{X}^I - X^{I\prime}) - \frac{i}{2} e^{im(\tau+\sigma)} (\dot{X}^I + X^{I\prime}) \\ &= -ie^{im\tau} \cos m\sigma \, \dot{X}^I + e^{im\tau} \sin m\sigma \, X^{I\prime} . \end{aligned} \tag{12.135}$$

This equation has taken the form

$$[L_m^\perp , X^I(\tau,\sigma)] = \xi^\tau_m \, \dot{X}^I + \xi^\sigma_m \, X^{I\prime}, \tag{12.136}$$

where

$$\begin{aligned}\xi^\tau_m(\tau,\sigma) &= -ie^{im\tau} \, \cos m\sigma \, , \\ \xi^\sigma_m(\tau,\sigma) &= \quad e^{im\tau} \, \sin m\sigma \, . \end{aligned} \tag{12.137}$$

The interpretation of (12.136), we claim, is that the Virasoro operators generate reparameterizations of the world-sheet. In particular, they change the τ and σ coordinates as

$$\begin{aligned}\tau &\longrightarrow \tau + \epsilon \, \xi^\tau_m(\tau,\sigma) \, , \\ \sigma &\longrightarrow \sigma + \epsilon \, \xi^\sigma_m(\tau,\sigma) \, , \end{aligned} \tag{12.138}$$

where ϵ is an infinitesimal parameter. In order to see this, note that Taylor expansion gives

$$\begin{aligned} X^I\big(\tau + \epsilon \, \xi^\tau_m \, , \sigma + \epsilon \, \xi^\sigma_m\big) &= X^I(\tau,\sigma) + \epsilon \, (\xi^\tau_m \, \dot{X}^I + \xi^\sigma_m \, X^{I\prime}) \\ &= X^I(\tau,\sigma) + \epsilon \Big[\, L_m^\perp , X^I(\tau,\sigma) \Big] . \end{aligned} \tag{12.139}$$

This equation states that the action of the Virasoro operators on the string coordinates generates the same change that would occur as a result of a reparameterization of the world-sheet. This is what we wanted to show.

What is the reparameterization generated by $L_0^\perp$? Setting $m = 0$ in (12.137), we find $\xi_0^\tau = -i$ and $\xi_0^\sigma = 0$. As a result, (12.136) gives

$$[L_0^\perp, X^I] = -i\,\partial_\tau X^I\,, \tag{12.140}$$

which we recognize as the Heisenberg equation of motion for X^I. Indeed, $L_0^\perp$ is, up to an additive constant, the string Hamiltonian, and as such, it must generate time translations. It is also interesting to note that for all m, ξ_m^σ vanishes at $\sigma = 0$ and at $\sigma = \pi$. This means that the reparameterizations generated by the Virasoro operators do not change the σ coordinates of the string endpoints. The range of σ remains $[0, \pi]$.

The functions ξ_m^τ and ξ_m^σ in (12.137) are not real, and used in (12.138) they spoil the reality of the coordinates τ and σ. This complication is familiar to you from quantum mechanics. Real transformations are generated by anti-Hermitian operators. The momentum operator $\vec{p} = -i\nabla$ is Hermitian, and therefore it is the anti-Hermitian combination $i\,\vec{p} = \nabla$ that generates real translations. Out of the operators $L_m^\perp$ and $L_{-m}^\perp$, we can generate two anti-Hermitian combinations:

$$L_m^\perp - L_{-m}^\perp \quad \text{and} \quad i\,(L_m^\perp + L_{-m}^\perp)\,. \tag{12.141}$$

Consider the first combination. It follows from (12.136) that the parameters for the transformation generated by $(L_m^\perp - L_{-m}^\perp)$ are

$$\begin{aligned} \xi^\tau &= \xi_m^\tau - \xi_{-m}^\tau = 2\sin m\tau\,\cos m\sigma\,, \\ \xi^\sigma &= \xi_m^\sigma - \xi_{-m}^\sigma = 2\cos m\tau\,\sin m\sigma\,. \end{aligned} \tag{12.142}$$

Quick calculation 12.6 Show that the parameters for the transformation generated by $i\,(L_m^\perp + L_{-m}^\perp)$ are

$$\begin{aligned} \xi^\tau &= \;\; 2\cos m\tau\,\cos m\sigma, \\ \xi^\sigma &= -2\sin m\tau\,\sin m\sigma. \end{aligned} \tag{12.143}$$

Our discussion of the Virasoro operators has been quite detailed. We have examined their precise definition, and we have seen how they affect the computation of the mass. We have determined their commutator algebra, and we have exhibited how they act on string coordinates. In the following section we will see that the Virasoro operators also enter into the definition of the operators that generate Lorentz transformations.

12.5 Lorentz generators

In Chapter 8, the Lorentz invariance of the string action allowed us to find a set of conserved world-sheet currents $\mathcal{M}^\alpha_{\mu\nu}$ labeled by the indices μ and ν, with $\mu \neq \nu$. The resulting conserved charges $M_{\mu\nu}$ were given in (8.65), and for open strings with $\sigma \in [0, \pi]$ they read

$$M_{\mu\nu} = \int_0^\pi \mathcal{M}^\tau_{\mu\nu}(\tau, \sigma)\,d\sigma = \int_0^\pi (X_\mu \mathcal{P}^\tau_\nu - X_\nu \mathcal{P}^\tau_\mu)\,d\sigma. \tag{12.144}$$

Making use of (12.1) and raising the spacetime indices,

$$M^{\mu\nu} = \frac{1}{2\pi\alpha'} \int_0^\pi \left(X^\mu \dot{X}^\nu - X^\nu \dot{X}^\mu \right) d\sigma. \tag{12.145}$$

Constructing suitable quantum operators can be delicate, so let us gain some intuition by thinking classically. Explicit mode expansions for X^μ and $\dot{X}^\nu$ are given in equations (9.56) and (9.57). Since $M^{\mu\nu}$ is guaranteed to be τ-independent, to evaluate (12.145) it suffices to pick up the τ-independent terms that arise in the products. For example,

$$X^\mu \dot{X}^\nu = x_0^\mu (\sqrt{2\alpha'}\,\alpha_0^\nu) + i\,2\alpha' \sum_{n\neq 0} \frac{1}{n}\, \alpha_n^\mu \alpha_{-n}^\nu \cos^2 n\sigma + \cdots, \tag{12.146}$$

where the dots represent τ-dependent terms that must fail to contribute in the calculation of $M^{\mu\nu}$. With this equation, and a similar one with μ and ν exchanged, we find that upon integration (12.145) gives

$$M^{\mu\nu} = x_0^\mu\, p^\nu - x_0^\nu\, p^\mu - i \sum_{n=1}^{\infty} \frac{1}{n} \left(\alpha_{-n}^\mu \alpha_n^\nu - \alpha_{-n}^\nu \alpha_n^\mu \right). \tag{12.147}$$

Quick calculation 12.7 Prove equation (12.147).

Equation (12.147) gives the classical Lorentz generators in terms of oscillation modes. We should ask ourselves whether we can use it, with the α recognized as operators, to define the quantum Lorentz generators. We will use (12.147) to *suggest* the form of the quantum Lorentz generators in light-cone gauge string theory. Since the canonical structure of the theory in the light-cone gauge is unusual, there is no guarantee that we can build consistent quantum Lorentz generators. An inability to construct quantum Lorentz generators would mean that quantum string theory fails to be physically Lorentz invariant.

In light-cone gauge, the most delicate quantum Lorentz generator is M^{-I} because the X^- coordinate is a rather nontrivial function of the transverse coordinates. A consistent M^{-I} must generate Lorentz transformations on the string coordinates, possibly accompanied by world-sheet reparameterizations. Indeed, in the simpler context of the point particle, the action of M^{-I} includes world-line reparameterizations. The generator M^{-I} must also satisfy the commutation relation

$$\left[M^{-I}, M^{-J} \right] = 0\,. \tag{12.148}$$

To find a candidate for M^{-I}, we consider equation (12.147) and write

$$M^{-I} \stackrel{?}{=} x_0^-\, p^I - x_0^I\, p^- - i \sum_{n=1}^{\infty} \frac{1}{n} \left(\alpha_{-n}^- \alpha_n^I - \alpha_{-n}^I \alpha_n^- \right). \tag{12.149}$$

This is just a first guess, though it is a pretty good one. A satisfactory M^{-I} should be both Hermitian and normal-ordered. Let us consider Hermiticity first. The first term in the right-hand side of (12.149) is Hermitian since x_0^- and p^I commute. The second term, however,

is not, since x_0^I and p^- do not commute. A simple solution is to symmetrize the term by writing

$$M^{-I} \stackrel{?}{=} x_0^- \, p^I - \frac{1}{2}\left(x_0^I \, p^- + p^- x_0^I\right) - i\sum_{n=1}^{\infty}\frac{1}{n}\left(\alpha_{-n}^-\alpha_n^I - \alpha_{-n}^I\alpha_n^-\right) . \tag{12.150}$$

The last term above is fully Hermitian since $(\alpha_n^I)^\dagger = \alpha_{-n}^I$ and $(\alpha_n^-)^\dagger = \alpha_{-n}^-$. Consider now normal ordering. Do all of the annihilation operators appear to the right of the creation operators? They do, because the α^- oscillators are normal-ordered Virasoro operators. Finally, to be complete, we must give the definition of p^-. As stated in (12.106), p^- includes an undetermined constant a that reflects our difficulties in ordering the Virasoro operator $L_0^\perp$. With this definition, and writing the other minus oscillators in terms of Virasoro operators, we find

$$\begin{aligned} M^{-I} = x_0^- p^I &- \frac{1}{4\alpha' p^+}\left(x_0^I\left(L_0^\perp + a\right) + \left(L_0^\perp + a\right)x_0^I\right) \\ &- \frac{i}{\sqrt{2\alpha'}\,p^+}\sum_{n=1}^{\infty}\frac{1}{n}\left(L_{-n}^\perp\alpha_n^I - \alpha_{-n}^I L_n^\perp\right) . \end{aligned} \tag{12.151}$$

Now that we have a candidate for the quantum Lorentz charge M^{-I}, we can discuss the computation of $[M^{-I}\,, M^{-J}]$.

There is much at stake in this calculation. It is in fact, one of the most important calculations in string theory. Our Lorentz charge has two undetermined parameters: the dimension D of spacetime, implicit in the sums over transverse directions, and the constant a affecting the mass of the particles. The calculation is long and uses many of our previously derived results, including the Virasoro commutation relations. We will not attempt to do it here, but the result is

$$\begin{aligned} \left[M^{-I}, M^{-J}\right] = &-\frac{1}{\alpha' p^{+2}}\sum_{m=1}^{\infty}\left(\alpha_{-m}^I\alpha_m^J - \alpha_{-m}^J\alpha_m^I\right) \\ &\times\left\{m\left[1 - \frac{1}{24}(D-2)\right] + \frac{1}{m}\left[\frac{1}{24}(D-2) + a\right]\right\} . \end{aligned} \tag{12.152}$$

The right-hand side is a sum of terms, each of which contains the operator $\left(\alpha_{-m}^I\alpha_m^J - \alpha_{-m}^J\alpha_m^I\right)$ for a different value of m. Such terms cannot cancel each other, so the commutator above vanishes if and only if the coefficient in large braces vanishes for all positive integers m:

$$m\left[1 - \frac{1}{24}(D-2)\right] + \frac{1}{m}\left[\frac{1}{24}(D-2) + a\right] = 0\,, \quad \forall\, m \in \mathbb{Z}^+ . \tag{12.153}$$

It suffices to examine this condition for $m = 1$ and $m = 2$ to conclude that each of the terms in brackets must simply vanish. We therefore have

$$1 - \frac{1}{24}(D-2) = 0 \quad \text{and} \quad \frac{1}{24}(D-2) + a = 0\,. \tag{12.154}$$

The first equation fixes the dimension of spacetime:

$$D = 26\,. \tag{12.155}$$

The second equation then fixes the constant a:

$$a = -\frac{1}{24}(D-2) = -\frac{24}{24} = -1\,. \tag{12.156}$$

This value of a coincides with the one obtained in (12.111) by ordering $L_0^\perp$ and using the zeta function to interpret the resulting infinity. For future reference, with $a = -1$ the expression for p^- in (12.106) becomes

$$2\alpha' p^- \equiv \frac{1}{p^+}\left(L_0^\perp - 1\right). \tag{12.157}$$

In addition, because of (12.14) the string Hamiltonian is now just

$$H = L_0^\perp - 1\,. \tag{12.158}$$

Here, of course, $L_0^\perp$ is the normal-ordered operator without additional constants. The above equation is the precise version of equation (12.16).

In summary, we have seen that the condition of Lorentz invariance of quantum string theory simultaneously fixes the dimension of spacetime and the constant shift in the masses of the particles. In superstring theory a similar calculation fixes the dimensionality of spacetime to the value $D = 10$. The fact that string theory cannot be a good Lorentz invariant quantum theory in any arbitrary dimension shows that string theory is very constrained. Even more, since the dimension of spacetime is uniquely selected by the requirement of consistency, we can say that string theory predicts the dimension of spacetime!

12.6 Constructing the state space

The classical open string does not provide a reasonable theory of physics because the mass of string states assumes a continuous range of values. Only the ground state is massless in the classical theory, and this ground state does not include any polarization labels. As a result, classical open strings have no states that can be identified as photons. The miracle of quantum string theory is that both of these problems are solved. The continuous spectrum disappears after quantization. Candidate photon states emerge because the downward shift of the squared masses gives us massless states with polarization labels.

Let us begin by introducing the ground states of the quantum string. The quantum string shares with the quantum point particle the same set of zero modes. We have the canonical pairs (x_0^I, p^I) and (x_0^-, p^+). Therefore, just as for the point particle (see (11.43)), we introduce states

$$|p^+,\ \vec{p}_T\rangle. \tag{12.159}$$

The above states are called *ground states* for all values of the momenta indicated by the labels. They are also declared to be vacuum states for all the oscillators in string theory. Thus, by definition, they are annihilated by all the a_n^I:

$$a_n^I|p^+, \vec{p}_T\rangle = 0\,, \quad n \geq 1, \quad I = 2, \ldots, 25. \tag{12.160}$$

How do we create states from the $|p^+,\ \vec{p}_T\rangle$? We simply act on them with the creation operators. There are infinitely many of them, and we can operate on each state arbitrarily many times with any particular creation operator. The list of creation operators at our disposal is infinite, but it can be organized as follows:

$$\begin{array}{cccc} a_1^{(2)\dagger} & a_1^{(3)\dagger} & \cdots & a_1^{(25)\dagger} \\ a_2^{(2)\dagger} & a_2^{(3)\dagger} & \cdots & a_2^{(25)\dagger} \\ \vdots & \vdots & \vdots & \vdots \\ a_n^{(2)\dagger} & a_n^{(3)\dagger} & \cdots & a_n^{(25)\dagger} \\ \vdots & \vdots & \vdots & \vdots \end{array} \tag{12.161}$$

Above, the polarization index I has been enclosed by parentheses. The general basis state $|\lambda\rangle$ of the state space can be written as

$$|\lambda\rangle = \prod_{n=1}^{\infty} \prod_{I=2}^{25} \left(a_n^{I\,\dagger}\right)^{\lambda_{n,I}} |p^+,\ \vec{p}_T\rangle\,. \tag{12.162}$$

Here the non-negative integer $\lambda_{n,I}$ denotes the number of times that the creation operator $a_n^{I\,\dagger}$ appears. As you can see, the state $|\lambda\rangle$ is specified by stating how many times each of the oscillators in the list (12.161) acts on the ground state. This information is given by the list of non-negative integers $\lambda_{n,I}$ for all $n \geq 1$ and all $I = 2, \ldots, 25$. Since all the creation operators commute among each other, the order in which they appear is irrelevant. We restrict ourselves to the case where states only have a finite number of creation operators acting on the ground states. This means that for each state $|\lambda\rangle$ only a finite number of $\lambda_{n,I}$ are different from zero. The string Hilbert space is an infinite-dimensional vector space: it is spanned by an infinite set of linearly independent basis states $|\lambda\rangle$. This is why string

theory describes an infinite number of different particles! A general state in the Hilbert space is a linear superposition of the basis states $|\lambda\rangle$.

To understand the physical significance of the above states, consider the mass-squared operator (12.108), with our new found knowledge that $a = -1$:

$$M^2 = \frac{1}{\alpha'}\Big(-1 + \sum_{n=1}^{\infty} n a_n^{I\,\dagger} a_n^I\Big)\,. \tag{12.163}$$

The sum appearing in (12.163) is important enough to have its own name; it is called the number operator $N^\perp$:

$$N^\perp \equiv \sum_{n=1}^{\infty} n a_n^{I\,\dagger} a_n^I\,, \quad M^2 = \frac{1}{\alpha'}(-1 + N^\perp)\,. \tag{12.164}$$

$N^\perp$ is the sum of standard number operators, one for each harmonic oscillator in the string. The main property of $N^\perp$ is that its commutator with a creation operator gives the mode number of that operator:

$$\big[N^\perp, a_n^{I\,\dagger}\big] = n\, a_n^{I\,\dagger}\,, \tag{12.165}$$

as you can readily verify. In addition,

$$\Big[N^\perp, a_n^I\Big] = -n\, a_n^I\,. \tag{12.166}$$

Since the number operator is normal-ordered it annihilates the ground states:

$$N^\perp |p^+,\ \vec{p}_T\rangle = 0\,. \tag{12.167}$$

Note, incidentally, that the number operator $N^\perp$ enters into the definition of $L_0^\perp$ in (12.105). We can write

$$L_0^\perp = \alpha'\, p^I p^I + N^\perp\,. \tag{12.168}$$

Let us get some practice using $N^\perp$ by computing its action on some basis states. Consider, for example, its action on $a_2^{I\,\dagger}|p^+,\ \vec{p}_T\rangle$:

$$N^\perp a_2^{I\,\dagger}|p^+,\ \vec{p}_T\rangle = \Big[N^\perp, a_2^{I\,\dagger}\Big]|p^+,\ \vec{p}_T\rangle + a_2^{I\,\dagger} N^\perp |p^+,\ \vec{p}_T\rangle = 2a_2^{I\,\dagger}|p^+,\ \vec{p}_T\rangle\,.$$

The state is an eigenstate of $N^\perp$ with eigenvalue 2. Now let us try a more complicated state:

$$\begin{aligned} N^\perp a_3^{J\,\dagger} a_2^{I\,\dagger}|p^+, \vec{p}_T\rangle &= \Big[N^\perp, a_3^{J\,\dagger}\Big] a_2^{I\,\dagger}|p^+, \vec{p}_T\rangle + a_3^{J\,\dagger} N^\perp a_2^{I\,\dagger}|p^+, \vec{p}_T\rangle \\ &= 5\, a_3^{J\,\dagger} a_2^{I\,\dagger}|p^+, \vec{p}_T\rangle\,. \end{aligned} \tag{12.169}$$

It is clear that when the number operator acts on a basis state, the eigenvalue is the sum of the mode numbers of the creation operators appearing in the state. In general, for the basis state $|\lambda\rangle$ in (12.162) we have

$$N^{\perp}|\lambda\rangle = N_{\lambda}^{\perp}|\lambda\rangle\,, \quad \text{with} \quad N_{\lambda}^{\perp} = \sum_{n=1}^{\infty}\sum_{I=2}^{25} n\lambda_{n,I}\,. \tag{12.170}$$

Since $N^{\perp}$ enters additively into the mass-squared operator (12.164), we see that the oscillator with mode number n contributes n units of $1/\alpha'$ to M^2. The eigenvalues of $N^{\perp}$ are non-negative integers, so for all string states $M^2 \geq -1/\alpha'$.

The open string state space is naturally equipped with an inner product. To define this inner product we introduce bras $\langle p^+, \vec{p}_T|$ which are the Hermitian conjugates of the kets $|p^+, \vec{p}_T\rangle$, and declare that

$$\langle p'^+, \vec{p}\,'_T|\, p^+, \vec{p}_T\rangle = \delta(p'^+ - p^+)\,\delta\,(\vec{p}\,'_T - \vec{p}_T)\,. \tag{12.171}$$

The delta functions here are necessary because the Hermiticity of the operators p^+ and $\vec{p}_T$ guarantees that the overlap above vanishes unless the p^+ and $\vec{p}_T$ eigenvalues of the states are identical. For the basis state $|\lambda\rangle$ in (12.162), we introduce the Hermitian conjugate bra $\langle\lambda|$ defined by

$$\langle\lambda| = \langle p^+,\ \vec{p}_T|\, \prod_{n=1}^{\infty}\prod_{I=2}^{25}\left(a_n^I\right)^{\lambda_{n,I}}. \tag{12.172}$$

The Hermitian conjugate of $|\lambda\rangle b$, where b is a number, is simply $b^*\langle\lambda|$. The inner product $(\lambda'\,,\lambda)$ between two basis states $|\lambda'\rangle$ and $|\lambda\rangle$ is defined as

$$(\lambda'\,,\lambda) = \langle\lambda'|\,\lambda\rangle\,. \tag{12.173}$$

For arbitrary states, the inner product is defined by declaring it to be linear on the second argument and antilinear on the first argument. To evaluate the overlap $\langle\lambda'|\lambda\rangle$ one moves the annihilation operators in $\langle\lambda'|$ towards the ground state in $|\lambda\rangle$ and the creation operators in $|\lambda\rangle$ towards the ground state in $\langle\lambda'|$. In moving these operators across each other we use the commutation relations (12.64), which appear in the standard form because in the overlap the annihilation operators are to the left of the creation operators. The commutators give $+1$ whenever they are nonvanishing, so the result of simplifying the oscillator content of the overlap is a positive number times the basic overlap in (12.171). For example, take $|\lambda'\rangle = a_1^{I\dagger}|p'^+, \vec{p}\,'_T\rangle$ and $|\lambda\rangle = a_1^{J\dagger}|p^+, \vec{p}_T\rangle$. We then find

$$\langle\lambda'|\,\lambda\rangle = \langle p'^+, \vec{p}\,'_T|\, a_1^I\, a_1^{J\dagger}|p^+, \vec{p}_T\rangle = \delta^{IJ}\delta(p'^+ - p^+)\,\vec{\delta}\,(\vec{p}\,'_T - \vec{p}_T)\,, \tag{12.174}$$

where we evaluated the overlap by replacing the product of oscillators by their commutator and using (12.171). Since the delta functions are positive functions, we conclude that all basis states $|\lambda\rangle$ have positive norm: $(\lambda,\lambda) > 0$. This inner product does not vanish because the annihilation operators on the bra match with the creation operators on the ket. Technically, (λ,λ) is infinite because of the delta functions, but this infinity becomes harmless when we consider continuous superpositions of states. Using the linearity and antilinearity of the inner product, we conclude that any state has positive norm. This is consistent with our statement that the open string state space is a Hilbert space.

Quick calculation 12.8 Explain why $(\lambda', \lambda) = 0$ whenever the basis states $|\lambda'\rangle$ and $|\lambda\rangle$ are different.

For each state $|\lambda\rangle$ in (12.162), we can construct the corresponding time-dependent physical state

$$\exp\left(-i(L_0^\perp - 1)\,\tau\right)|\lambda\rangle\,, \tag{12.175}$$

which satisfies the Schrödinger equation with Hamiltonian (12.158). In the light-cone quantization of the string a time-dependent state is physical if it satisfies the Schrödinger equation. We will consider more general superpositions of time-dependent states in the following section.

We are now ready to discuss particular states in some detail. We will begin with the simplest ones, the ground states. These are the unique states with $N^\perp = 0$. As in the case of the point particle, the states $|p^+, \vec{p}_T\rangle$ are the one-particle states of a scalar field. They are states of a scalar particle. What is the mass of this particle? To find out, we act on the states with M^2:

$$M^2|\,p^+,\ \vec{p}_T\rangle = \frac{1}{\alpha'}(-1 + N^\perp)|\,p^+,\ \vec{p}_T\rangle = -\frac{1}{\alpha'}\,|\,p^+,\ \vec{p}_T\rangle\,. \tag{12.176}$$

The value of M^2 is all due to the ordering constant. If this constant had vanished the mass would have been zero. In fact, massless scalars are problematic – they have not been observed in nature. The result, however, is strange: $M^2 = -1/\alpha' < 0$. The wavefunction of the state tells the same story: $\psi(\tau, p^+, \vec{p}_T)$ can be set in correspondence with a classical scalar field. This scalar field, which has a negative mass-squared, is called a *tachyon*. A negative mass-squared is a sign of instability: the potential for a scalar field goes like $V = \frac{1}{2}M^2\phi^2$ (see (10.2)), so a negative M^2 simply means that the stationary point $\phi = 0$ is unstable. The energy can be reduced by having $\phi \neq 0$. We will study this further in Section 12.8.

Let us consider now the excited states with lowest M^2. Those arise when $N^\perp$ takes the smallest possible nonzero value, $N^\perp = 1$. Remarkably, due to the ordering constant, the $N^\perp = 1$ states have $M^2 = 0$. They are massless states. Had the ordering constant taken a noninteger value, quantum string theory would have no massless states. We get states with $N^\perp = 1$ when we act with any of the transverse oscillators $a_1^{I\,\dagger}$ on the ground states $|p^+, \vec{p}_T\rangle$. That means that we have $D - 2 = 24$ massless states:

$$a_1^{I\,\dagger}|p^+,\ \vec{p}_T\rangle\,, \quad M^2 a_1^{I\,\dagger}|p^+,\ \vec{p}_T\rangle = 0\,. \tag{12.177}$$

The general massless state is a linear combination of the above basis states:

$$\sum_{I=2}^{25} \xi_I\, a_1^{I\,\dagger}|p^+,\ \vec{p}_T\rangle\,. \tag{12.178}$$

The above expression may remind you of the photon states (10.88) that we found in our light-cone analysis of Maxwell theory:

$$\sum_{I=2}^{D-1} \xi_I \, a^{I\dagger}_{p^+,\, p_T} |\Omega\rangle \,. \tag{12.179}$$

We have a matching of states: in both cases ξ_I is an arbitrary transverse vector, and the states correspond to one another:

$$a_1^{I\dagger} |p^+, \vec{p}_T\rangle \;\longleftrightarrow\; a^{I\dagger}_{p^+,\, p_T} |\Omega\rangle \,. \tag{12.180}$$

Both states have exactly the same Lorentz labels, they carry the same momenta, and have the same mass. This proves a remarkable result. The open string theory quantum states include photon states! Our study of open string theory started from the Nambu–Goto action. This action has no hint whatsoever of electromagnetic gauge invariance. Nevertheless, we have shown that open string theory contains Maxwell field excitations. This astonishing result is largely due to the mass shift encountered in passing from the classical to the quantum theory of the open string.

It is worth belaboring the point. In Chapter 10 we showed that the quantum states of free Maxwell theory – the photon states – are $(D-2)$ massless states, labeled by a transverse Lorentz index. The index is important: it indicates that these states transform into each other under Lorentz transformations. Exactly these kind of states have appeared in our quantization of the string. Additionally, the collection of wavefunctions $\psi_I(\tau, p^+, \vec{p}_T)$ associated with the states in (12.177) matches with the components A_I of the Maxwell gauge field. Finally, the Schrödinger equation for these wavefunctions matches the light-cone gauge field equation for the Maxwell field. We will show this in the following section.

Let us conclude with an examination of the states with $N^\perp = 2$. They are built by acting on the ground states with $a_1^{I\dagger} a_1^{J\dagger}$ or with $a_2^{I\dagger}$. The number of states built with $a_1^{I\dagger} a_1^{J\dagger}$ is the same as the number of independent entries in a symmetric matrix of size $(D-2)$, namely $\frac{1}{2}(D-2)(D-1)$. The number of states built with $a_2^{I\dagger}$ is $(D-2)$. The total number of states is therefore

$$\frac{1}{2}(D-2)(D-1) + (D-2) = \frac{1}{2}(D-2)(D+1)\,, \tag{12.181}$$

and their mass-squared is given by $M^2 = 1/\alpha'$. These particles are known as massive tensors, and in $D = 26$ there are 324 such states. Our results for all states with $N^\perp \leq 2$ are summarized in Table 12.1. A formula useful to count states at higher values of $N^\perp$ is derived in Problem 12.11. The formula tells you that at $N^\perp = 3$, for example, you get 2600 states by acting with $a_1^{I\dagger} a_1^{J\dagger} a_1^{K\dagger}$ on the ground states.

Each state $|\lambda\rangle$ of the quantum string represents a one-particle state of fixed momentum. Thus, the $a_1^{I\dagger}|p^+, \vec{p}_T\rangle$ are one-photon states, and the $a_1^{I\dagger} a_1^{J\dagger}|p^+, \vec{p}_T\rangle$ are one-particle tensor states (*not* two-photon states). Each state $|\lambda\rangle$ has discrete labels $\lambda_{n,I}$ and continuous labels p^+ and $\vec{p}_T$. There is one wavefunction for each set of discrete labels, as you can see in the table. Accordingly, there is one quantum field for each set of discrete labels. The multiparticle states are described using these quantum fields. The total quantum field

Table 12.1 List of open string states with $N^\perp \leq 2$

$N^\perp$	$\|\lambda\rangle$	$\alpha' M^2$	Number of states	Wavefunction
0	$\|p^+, \vec{p}_T\rangle$	-1	1	$\psi(\tau, p^+, \vec{p}_T)$
1	$a_1^{I\dagger}\|p^+, \vec{p}_T\rangle$	0	$D-2$	$\psi_I(\tau, p^+, \vec{p}_T)$
2	$a_1^{I\dagger}a_1^{J\dagger}\|p^+, \vec{p}_T\rangle$	1	$\frac{1}{2}(D-2)(D+1)$	$\psi_{IJ}(\tau, p^+, \vec{p}_T)$
	$a_2^{I\dagger}\|p^+, \vec{p}_T\rangle$			$\psi_I(\tau, p^+, \vec{p}_T)$

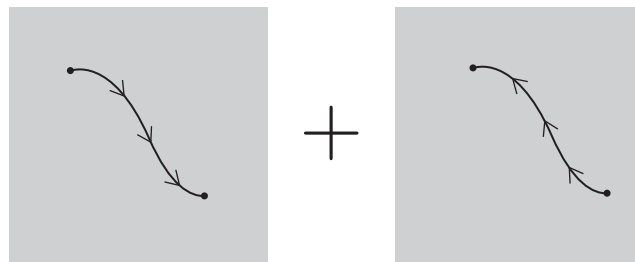

Fig. 12.1 **An unoriented open string is a quantum state that can be obtained by superposition of states that differ only by orientation. The unoriented state is invariant under orientation reversal.**

theory which describes the whole set of quantum fields associated with the one-particle states of the string is called *string field theory*.

The quantum theory we have discussed so far is that of *oriented* open strings. The quantum operator $X^I(\tau, \sigma)$ involves a parameter $\sigma \in [0, \pi]$ and thus an orientation, defined as the direction of increasing σ. It is possible to define a theory of *unoriented* strings (Problem 12.12). The key idea is that one can define an operator Ω that is a symmetry of the theory (commutes with the Hamiltonian) and reverses the orientation of strings. The theory of unoriented strings is obtained by restricting the oriented string spectrum to the set of states that are invariant under the action of Ω. Unoriented strings are not strings without orientation: they should be viewed as quantum superposition of states that, as a whole, are invariant under orientation reversal. We can imagine an unoriented state as the superposition of a string state and the same state with opposite orientation, as shown in Figure 12.1.

12.7 Equations of motion

To elaborate on the correspondence between string states and quantum fields we now consider the Schrödinger equations satisfied by the string wavefunctions. We saw in Section 11.4 that the Schrödinger equation for the point particle wavefunction is isomorphic to the classical field equation of a scalar field. We want to repeat such an analysis for the string.

To construct general time-dependent states from the string basis states we need wavefunctions. Consider, for example, a basis state

$$a_{n_1}^{I_1\,\dagger} \ldots a_{n_k}^{I_k\,\dagger} |p^+, \vec{p}_T\rangle \,. \tag{12.182}$$

The general time-dependent state built by superposition is

$$|\Psi, \tau\rangle = \int dp^+ \, d\vec{p}_T \; \psi_{I_1 \ldots I_k}(\tau, p^+, \vec{p}_T) \; a_{n_1}^{I_1\,\dagger} \ldots a_{n_k}^{I_k\,\dagger} |p^+, \vec{p}_T\rangle \,. \tag{12.183}$$

The polarization indices carried by the oscillators match with the index labels of the wavefunction $\psi_{I_1 \ldots I_k}(\tau, p^+, \vec{p}_T)$. This equation is the string analog of (11.48), which gives the general time-dependent state of the point particle. For general tachyon states (12.183) becomes

$$|\text{tachyon}, \tau\rangle = \int dp^+ \, d\vec{p}_T \; \psi(\tau, p^+, \vec{p}_T) \, |p^+, \vec{p}_T\rangle \,. \tag{12.184}$$

For photon states we write

$$|\text{photon}, \tau\rangle = \int dp^+ \, d\vec{p}_T \; \psi_I(\tau, p^+, \vec{p}_T) \; a_1^{I\,\dagger} |p^+, \vec{p}_T\rangle \,. \tag{12.185}$$

The Schrödinger equation satisfied by the general states (12.183) is

$$i \frac{\partial}{\partial \tau} |\Psi, \tau\rangle = H |\Psi, \tau\rangle \,. \tag{12.186}$$

Here, the Hamiltonian is given by

$$H = (L_0^\perp - 1) = \alpha' p^I p^I + N^\perp - 1 = \alpha' (p^I p^I + M^2) \,, \tag{12.187}$$

where we used (12.158) and (12.168). Using the explicit expression (12.183) for the states, equation (12.186) gives:

$$i \frac{\partial}{\partial \tau} \psi_{I_1 \ldots I_k} = \left(\alpha' p^I p^I + N^\perp - 1 \right) \psi_{I_1 \ldots I_k} \,, \tag{12.188}$$

where $N^\perp$ denotes the eigenvalue of the operator $N^\perp$ for the state (12.183).

Quick calculation 12.9 Show that equation (12.188) emerges from the Schrödinger equation (12.186). The calculation parallels that which gave (11.53).

For the tachyon states (12.184), $N^\perp = 0$, and we get

$$i \frac{\partial \psi}{\partial \tau} = \left(\alpha' p^I p^I - 1 \right) \psi \,. \tag{12.189}$$

For the photon states (12.185), $N^\perp = 1$, and we get

$$i \frac{\partial \psi_I}{\partial \tau} = \alpha' p^J p^J \, \psi_I \,. \tag{12.190}$$

Let us now compare these Schrödinger equations with the relevant classical field equations. We showed in Chapter 10 that the scalar field equation

$$\left(\partial^2 - m^2 \right) \phi = 0 \,, \tag{12.191}$$

could be written as (10.30):

$$\left(i\,\frac{\partial}{\partial x^+} - \frac{1}{2p^+}(p^I p^I + m^2)\right)\phi(x^+, p^+, \vec{p}_T) = 0\,. \tag{12.192}$$

Letting $x^+ = 2\alpha' p^+ \tau$ we now have

$$\left(i\frac{\partial}{\partial \tau} - \alpha'(p^I p^I + m^2)\right)\phi(\tau, p^+, \vec{p}_T) = 0\,. \tag{12.193}$$

This equation is precisely the same as (12.189) when $m^2 = -1/\alpha'$, confirming the identification of the tachyon with a scalar field. Perhaps more surprisingly, equation (12.193) is structurally equivalent to the Schrödinger equation (12.188) satisfied by *any* string wavefunction. The only difference is that the wavefunctions carry indices. As a result, if the correspondence is to hold, the classical field equation for the field associated with any string state must take the form (12.191), with the field carrying some indices.

This may seem strange: is not the Maxwell classical field equation, for example, more complicated than the field equation for a scalar? Not in the light-cone gauge. We noticed this before: equation (10.83) showed that the transverse components of the gauge field satisfy $p^2 A^I(p) = 0$. This is of the form (12.191) with $m^2 = 0$. The steps leading from (12.191) to (12.193), when applied to $\partial^2 A^I = 0$ give

$$\left(i\frac{\partial}{\partial \tau} - \alpha' p^J p^J\right) A^I(\tau, p^+, \vec{p}_T) = 0\,. \tag{12.194}$$

This classical field equation for the Maxwell field is in complete correspondence with the Schrödinger equation (12.190) for the $N^\perp = 1$ wavefunctions.

12.8 Tachyons and D-brane decay

We conclude this chapter by discussing the physics of the tachyon. We explained earlier that the tachyon state has the lowest value of M^2:

$$M^2|p^+, \vec{p}_T\rangle = -\frac{1}{\alpha'}|p^+, \vec{p}_T\rangle\,. \tag{12.195}$$

The field associated with this state is a scalar field. What does it mean for this scalar to have a negative M^2? The physics of the open string tachyon was a mystery ever since the discovery of string theory. A series of developments starting in 1999 have essentially elucidated the role of the open string tachyon. Let us discuss what has been learned.

Our first goal is to understand the instability of a theory with a tachyon. For this purpose we consider the Lagrangian density for a classical scalar field, along the lines of Section 10.2. In some generality,

$$\mathcal{L} = -\frac{1}{2}\,\eta^{\mu\nu}\,\partial_\mu\phi\,\partial_\nu\phi - V(\phi) = \frac{1}{2}(\partial_0\phi)^2 - \frac{1}{2}|\nabla\phi|^2 - V(\phi)\,, \tag{12.196}$$

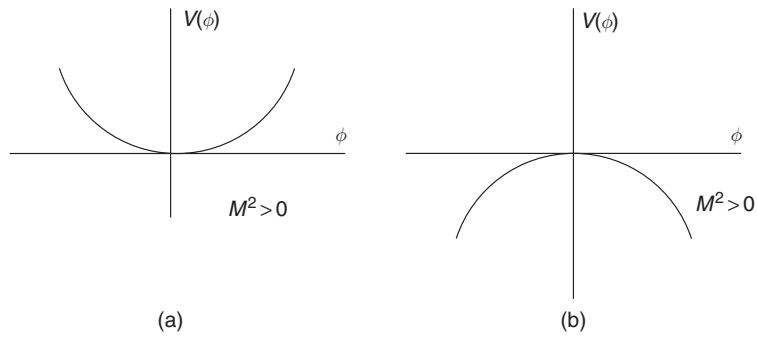

Fig. 12.2 **(a) A potential $V(\phi) = \frac{1}{2}M^2\phi^2$ with positive mass-squared M^2. The value $\phi = 0$ is a stable critical point. (b) A potential $V(\phi) = \frac{1}{2}M^2\phi^2$ with negative mass-squared M^2. The value $\phi = 0$ is an unstable critical point.**

where $V(\phi)$ is the *potential* for the scalar field. For spatially homogeneous field configurations, $\nabla\phi = 0$, and the potential energy density is given by the potential $V(\phi)$. The equation of motion following from variation is

$$\partial^2\phi - V'(\phi) = 0\,, \tag{12.197}$$

where prime denotes the derivative with respect to the argument. More explicitly,

$$-\frac{\partial^2\phi}{\partial t^2} + \nabla^2\phi - V'(\phi) = 0\,. \tag{12.198}$$

Quick calculation 12.10 Prove that equation (12.197) arises from variation of the action $S = \int d^D x\mathcal{L}$.

To understand the instability of the tachyon scalar field theory it suffices to consider the free part of the tachyon Lagrangian; interactions will feature later. For a free scalar field theory, the potential $V(\phi)$ takes the form

$$V(\phi) = \frac{1}{2}M^2\phi^2\,. \tag{12.199}$$

Here M^2 is the mass-squared of the scalar field (this potential will change with the inclusion of interactions). When $M^2 > 0$, the potential $V(\phi)$ has a stable minimum at $\phi = 0$. When $M^2 < 0$, $V(\phi)$ has an unstable maximum at $\phi = 0$ (Figure 12.2). We can understand the implications of such potentials by studying the equation of motion for the field. Using the specified form of V, equation (12.198) gives

$$-\frac{\partial^2\phi}{\partial t^2} + \nabla^2\phi - M^2\phi = 0\,. \tag{12.200}$$

To make our analysis simpler, let us assume that the field ϕ depends only on time. The equation of motion then becomes

$$\frac{d^2\phi(t)}{dt^2} + M^2\phi(t) = 0\,. \tag{12.201}$$

When $M^2 = M \cdot M > 0$, the solutions of this equation represent oscillations:

$$\phi = A\cos(Mt) + B\sin(Mt) = C\sin(Mt + \alpha_0)\,. \tag{12.202}$$

This is the interpretation of a scalar field with a "good" mass-squared. The scalar field could sit at $\phi = 0$ forever because it is a stable point; if it is displaced, it simply oscillates around $\phi = 0$.

Consider, on the other hand, the tachyon, which is an example of a scalar with negative mass-squared. In this case it is convenient to write $M^2 = -\beta^2 = -\beta \cdot \beta$, and equation (12.201) becomes

$$\frac{d^2\phi(t)}{dt^2} - \beta^2\phi(t) = 0\,, \tag{12.203}$$

with $\beta^2 > 0$. This time the solutions are

$$\phi(t) = A\cosh(\beta t) + B\sinh(\beta t)\,. \tag{12.204}$$

Consider the solution $\phi(t) = \sinh(\beta t)$. At time zero ϕ is zero, but as time goes to infinity ϕ also goes to infinity. We can imagine this as the field rolling to the right of the potential in Figure 12.2(b). In fact, for any nontrivial solution, ϕ necessarily reaches infinite absolute value either in the far past or in the far future. The tachyon could stay at $\phi = 0$ forever, using the trivial solution $\phi(t) = 0$, but any infinitesimal perturbation would set it on a course to a dramatic roll-off. The value $\phi = 0$ is an allowed critical point, but it is unstable. We cannot realistically expect the tachyon to stay near $\phi = 0$ for an indefinite length of time. This is the instability of a theory that contains a tachyon. Since the mass-squared of the open string tachyon is equal to $(-1/\alpha')$, the free part of the tachyon potential is

$$V_{\text{tach}}^{\text{free}}(\phi) = -\frac{1}{2\alpha'}\phi^2\,. \tag{12.205}$$

A mechanical analogy works for arbitrary potentials $V(\phi)$. You can visualize the spatially homogeneous rolling of a scalar field on a potential $V(\phi)$ by considering the motion of a particle on the potential $V(x)$, where x, replacing ϕ, is the coordinate along the motion. Indeed, the relevant equations match. For an arbitrary potential $V(\phi)$, homogeneous rolling is governed by

$$\frac{d^2\phi}{dt^2} = -V'(\phi)\,, \tag{12.206}$$

while the rolling of a unit-mass particle on a potential $V(x)$ is governed by Newton's second law:

$$\frac{d^2x}{dt^2} = -V'(x)\,. \tag{12.207}$$

The presence of a tachyon signals an instability of open string theory. More precisely, there is some instability in the theory of open strings on the background of a space-filling D25-brane. It is clear that we should try to understand the fate of this instability: once the tachyon begins to roll, where does it end? For a while, not everyone agreed that this was an urgent question. Some argued that along with the lack of fermions, the tachyon was another

good reason to consider this open string theory unrealistic and not worth much study. Some even saw the tachyon as a sign that open bosonic string theory is simply inconsistent. For quite a few years, superstring theories, the kind of string theories that also include fermions, seemed blessedly devoid of tachyons. Later studies, however, showed that tachyons can appear when we construct realistic models based on superstrings. It then became clear that we must try to understand tachyons.

The open string theory we have in our hands is the theory of strings on a D25-brane, a D-brane that fills all of the space dimensions. The D25-brane is a physical object, not just a mathematical construct, so it has a constant energy density T_{25} which, in fact, can be calculated exactly. The key insight can now be stated: the theory of open strings is, in some sense, the theory of the D25-brane itself! We have viewed tachyons as states of strings attached to a D-brane. A D-brane with open strings attached, it turns out, is an excited state of the D-brane. If this is so, a tachyon state represents an excitation that can lower the energy of the D-brane. The existence of the tachyon is telling us that the D25-brane is unstable!

Since the tachyon describes the physics of the D25-brane, the energy density of this brane is a contribution to the potential energy of the system, and it must be incorporated into the tachyon potential. As a result, the potential in (12.205) is changed into

$$V_{\text{tach}}(\phi) = T_{25} - \frac{1}{2\alpha'}\phi^2 + \beta\phi^3 + \cdots. \tag{12.208}$$

We have also included in the tachyon potential a cubic term and represented other possible terms by dots. All the terms that are cubic or higher order in the field represent the effect of interactions. The above potential describes correctly our statements about the D25-brane. The unstable point $\phi = 0$ represents the world with a D25-brane, and therefore has an energy density T_{25}. To find out what happens when the tachyon starts rolling down, we need to calculate the full tachyon potential $V_{\text{tach}}(\phi)$.

The physics can be anticipated before computing this potential. If the D25-brane is unstable, it will decay. The stable endpoint of this process would be a world without the D25-brane. If this is so, the tachyon potential must have a stable critical point at some $\phi = \phi^*$ with $V_{\text{tach}}(\phi^*) = 0$. That stable critical point would represent a background with zero energy, a background where the D25-brane, rendered unstable by the tachyon, has disappeared completely! The expected form of the tachyon potential is shown in Figure 12.3.

This proposal has now been verified convincingly. The complete tachyon potential was calculated using the field theory of open strings and, remarkably, it was possible to display a critical point with zero energy. With this result, and additional evidence obtained by other means, physicists have demonstrated that the tachyon instability is the instability of the D25-brane.

What happens when the tachyon rolls down to the stable minimum and the D25-brane disappears? All the open strings must also disappear, because open string endpoints are confined to D-branes. Closed strings, however, can exist in the absence of D-branes. All

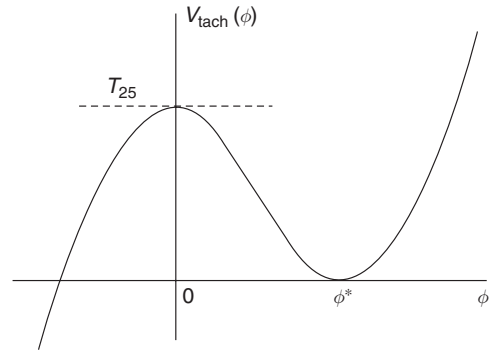

Fig. 12.3 **The tachyon potential for the open string theory based on a D25-brane. The configuration $\phi = 0$ represents the unstable D-brane. The stable critical point ϕ^* has zero energy.**

the energy initially stored in the D25-brane goes into closed strings. At the stable critical point ϕ^* all particles that arise as open string excitations, including the tachyon, must disappear. This shows that the theory near ϕ^* is quite subtle. *Vacuum string field theory* is an attempt to formulate string theory at the vacuum ϕ^* where both D-branes and open strings disappear.

Further interesting facts about tachyons and D-branes have emerged. It has been shown that Dp-branes with $p < 25$ are themselves large and coherent excitations of the tachyon field of the D25-brane. In some sense, D-branes are made of tachyons! This is also true, with minor modifications, in superstring theory. In superstring theory certain D-branes carry charge and therefore charge conservation ensures that they are stable against decay. In fact, the open string theory in the background of any such D-brane has no tachyons. However, a configuration consisting of a D-brane and a coincident, oppositely charged anti-D-brane, is unstable: the two objects can annihilate without violating charge conservation. The open strings which stretch from the D-brane to the anti-D-brane contain a tachyon – a superstring tachyon! This tachyon describes the instability of the D-brane/anti-D-brane pair. The study of D-brane/anti-D-brane annihilation plays an important role in attempts to use string theory to describe the early universe. It is thus possible that the tachyon will end up playing a prominent role in string cosmology.

Problems

Problem 12.1 Heisenberg equation for the momentum density.

We verified in (12.21) that $\dot{X}^I = 2\pi\alpha'\mathcal{P}^{\tau I}$ follows from the Heisenberg equation $i\dot{\xi} = [\xi, H]$, when $\xi = X^I$. Calculate $\dot{\mathcal{P}}^{\tau I}$, and use the result to verify that the classical equation of motion $\ddot{X}^I - X^{I\prime\prime} = 0$ holds as an operator equation.

Problem 12.2 Testing explicitly some vanishing commutators.

Use the mode expansion (12.32) and the commutation relations of the α to check explicitly that equation (12.26) holds: $[X^{I\,\prime}(\tau,\sigma),\, X^{J\,\prime}(\tau,\sigma')] = [\dot{X}^I(\tau,\sigma)\,,\, \dot{X}^J(\tau,\sigma')] = 0$.

Problem 12.3 Testing explicitly the main commutator.

(a) Use the explicit mode expansions of X^I and $\mathcal{P}^{\tau J}$, together with the commutation relations (12.45) and (12.54), to show that

$$\Big[\, X^I(\tau,\sigma)\,,\, \mathcal{P}^{\tau J}(\tau,\sigma')\,\Big] = i\,\eta^{IJ}\;\frac{1}{\pi}\sum_{n\in\mathbb{Z}}\cos n\sigma\;\cos n\sigma'\,.$$

(b) If the above result agrees with (12.10), we must have

$$\delta(\sigma-\sigma') = \frac{1}{\pi}\sum_{n\in\mathbb{Z}}\cos n\sigma\;\cos n\sigma'\,. \tag{1}$$

This equation follows from the completeness of the functions $\cos n\sigma$ with $n \geq 0$ on the interval $\sigma \in [0,\pi]$. The completeness is readily explained: any function $f(\sigma)$ defined over $\sigma \in [0,\pi]$ can be extended to a function over $\sigma \in [-\pi,\pi]$ by letting $f(-\sigma) \equiv f(\sigma)$ for $\sigma \in [0,\pi]$. The resulting function is an even function of σ and by the basic result of Fourier series it can be expanded in terms of cosines. We can therefore expand any function $f(\sigma)$ with $\sigma \in [0,\pi]$ as

$$f(\sigma) = \sum_{n=0}^{\infty} A_n \cos n\sigma\,. \tag{2}$$

Prove (1) by calculating A_n and substituting the result back into the right-hand side of (2).

Problem 12.4 Analytic continuation of the zeta function.

Consider the definition of the gamma function $\Gamma(s) = \int_0^\infty dt\, e^{-t}t^{s-1}$. Let $t \to nt$ in this integral, and use the resulting equation to prove that

$$\Gamma(s)\,\zeta(s) = \int_0^\infty dt\,\frac{t^{s-1}}{e^t-1}\,, \quad \Re(s) > 1\,. \tag{1}$$

Verify also the small t expansion

$$\frac{1}{e^t-1} = \frac{1}{t} - \frac{1}{2} + \frac{t}{12} + \mathcal{O}(t^2)\,. \tag{2}$$

Use the above equations to show that for $\Re(s) > 1$

$$\Gamma(s)\,\zeta(s) = \int_0^1 dt\, t^{s-1}\Big(\frac{1}{e^t-1} - \frac{1}{t} + \frac{1}{2} - \frac{t}{12}\Big) + \frac{1}{s-1} - \frac{1}{2s} + \frac{1}{12\,(s+1)} + \int_1^\infty dt\,\frac{t^{s-1}}{e^t-1}\,.$$

Explain why the right-hand side above is well defined for $\Re(s) > -2$. It follows that this right-hand side defines the analytic continuation of the left-hand side to $\Re(s) > -2$. Recall the pole structure of $\Gamma(s)$ (Problem 3.6) and use it to show that $\zeta(0) = -1/2$ and that $\zeta(-1) = -1/12$.

Problem 12.5 The Virasoro algebra is a Lie algebra.

A vector space L with elements $x, y, z, \ldots$ and a bilinear bracket $[\cdot\,,\,\cdot]$, that takes two elements of L and yields another element of L, is a Lie algebra if the two following conditions hold.

(i) Antisymmetry: $[x, y] = -[y, x]$ for all elements x and y of L.
(ii) Jacobi identity: $[x, [y, z]] + [y, [z, x]] + [z, [x, y]] = 0$, for all elements x, y, and z of L.

Consider the vector space L spanned by the Virasoro operators with modes $n \in \mathbb{Z}$. Show that the commutators in (12.124), assumed to hold for all values of m and n, define a Lie algebra. Then, consider the commutators in (12.133), and show that they also define a Lie algebra.

Problem 12.6 Consistency conditions on the Virasoro algebra central terms.

The Virasoro commutation relations take the form

$$[L_m^\perp, L_n^\perp] = (m-n)L_{m+n}^\perp + C(m)\delta_{m+n,0}\,, \tag{1}$$

where $C(m)$ is a function of m that was calculated directly in this chapter. The purpose of this problem is to find the constraints on $C(m)$ that follow from the condition that (1) defines a Lie algebra.

(a) What does the antisymmetry requirement on a Lie algebra tell you about $C(m)$? What is $C(0)$?
(b) Consider now the Jacobi identity for generators $L_m^\perp$, $L_n^\perp$, and $L_k^\perp$ with $m + n + k = 0$. Show that

$$(m-n)C(k) + (n-k)C(m) + (k-m)C(n) = 0\,. \tag{2}$$

(c) Use equation (2) to show that $C(m) = \alpha m$ and $C(m) = \beta m^3$, for constants α and β, yield consistent central extensions.
(d) Consider equation (2) with $k = 1$. Show that $C(1)$ and $C(2)$ determine all $C(n)$.

Problem 12.7 Exercises with Virasoro operators.

(a) Use the Virasoro algebra (12.133) to show that if a state is annihilated by L_1 and L_2 it is annihilated by all L_n with $n \geq 1$.
(b) Consider the Virasoro operators L_0, L_1, and L_{-1}. Write out the three relevant commutators. Do these operators form a subalgebra of the Virasoro algebra? Is there a central term in here? Calculate the result of acting with each of these three operators on the zero-momentum vacuum state $|0\rangle$.

Problem 12.8 Virasoro operators acting on states.

(a) Use (12.100) to write the state $L^{\perp}_{-6}|0\rangle$ as a finite sum of terms with oscillators acting on the vacuum. Do the same for $L^{\perp}_{-2}L^{\perp}_{-2}|0\rangle$.

(b) Write $L^{\perp}_{2}L^{\perp}_{-2}|0\rangle$ and $L^{\perp}_{-2}L^{\perp}_{2}|0\rangle$ as finite sums of terms with oscillators acting on the vacuum. Use the results to evaluate $[L^{\perp}_{2}, L^{\perp}_{-2}]|0\rangle$. Confirm that your evaluation is consistent with the Virasoro algebra (12.133).

Problem 12.9 Reparameterizations generated by Virasoro operators.

(a) Consider the string at $\tau = 0$. Which of the combinations in (12.141) reparameterizes the σ coordinate of the string while keeping $\tau = 0$? When $\tau = 0$ is preserved, the world-sheet reparameterization is actually a *string* reparameterization. Show that the generators of these reparameterizations form a subalgebra of the Virasoro algebra.

(b) Describe the general *world-sheet* reparameterization that leaves the midpoint $\sigma = \pi/2$ of the $\tau = 0$ open string fixed. Express this reparameterization using an infinite set of constrained parameters.

Problem 12.10 Reparameterizations and constraints.

(a) Verify that the reparameterization parameters in (12.137) satisfy the relations (omitting the subscript m for convenience)

$$\dot{\xi}^{\tau} = \xi^{\sigma\prime}\,, \quad \dot{\xi}^{\sigma} = \xi^{\tau\prime}\,.$$

(b) Think of the reparameterizations (12.138) generated by the Virasoro operators as a change of coordinates

$$\tau' = \tau + \epsilon\,\xi^{\tau}(\tau,\sigma)\,, \quad \sigma' = \sigma + \epsilon\,\xi^{\sigma}(\tau,\sigma)\,.$$

Note that for infinitesimal ϵ the above equations also imply that

$$\tau = \tau' - \epsilon\,\xi^{\tau}(\tau',\sigma')\,, \quad \sigma = \sigma' - \epsilon\,\xi^{\sigma}(\tau',\sigma')\,.$$

Show that the classical constraints

$$\partial_\tau X\cdot\partial_\sigma X = 0\,, \quad (\partial_\tau X)^2 + (\partial_\sigma X)^2 = 0\,,$$

assumed to hold in (τ,σ) coordinates, also hold in (τ',σ') coordinates (to order ϵ).

Problem 12.11 Counting symmetric products.[†]

Prove that the number N of different products of the form

$$a^{I_1}a^{I_2}\ldots a^{I_n}\,,$$

where the superscripts $I_1, I_2, \ldots, I_n$ run over k values $1, 2, \ldots, k$, is given by

$$N = \frac{(n+k-1)!}{n!\,(k-1)!} = \frac{k}{1}\cdot\frac{k+1}{2}\cdot\frac{k+2}{3}\cdots\frac{k+n-1}{n}\,.$$

Hint: one can represent each product using n identical balls and $k-1$ identical dividers. For $k=6$ and $n=9$, for example, we have:

$\longleftrightarrow \quad a^1\ a^2a^2\ a^4a^4a^4\ a^5\ a^6a^6\,.$

Problem 12.12 Unoriented open strings.

The open string $X^\mu(\tau,\sigma)$, with $\sigma\in[0,\pi]$ and fixed τ, is a parameterized curve in spacetime. The orientation of a string is the direction of increasing σ on this curve.

(a) Consider now the open string $X^\mu(\tau,\pi-\sigma)$ at the same τ. How is this second string related to the first string above? How are their endpoints and orientations related? Make a rough sketch showing the original string as a continuous curve in spacetime, and the second string as a dashed curve in spacetime.

Introduce an orientation reversing *twist* operator Ω such that

$$\Omega\, X^I(\tau,\sigma)\,\Omega^{-1} = X^I(\tau,\pi-\sigma)\,. \tag{1}$$

Moreover, declare that

$$\Omega\, x_0^-\,\Omega^{-1} = x_0^-\,, \quad \Omega\, p^+\,\Omega^{-1} = p^+\,. \tag{2}$$

(b) Use the open string oscillator expansion (12.32) to calculate

$$\Omega\, x_0^I\,\Omega^{-1}\,, \quad \Omega\,\alpha_0^I\,\Omega^{-1}\,, \quad \text{and} \quad \Omega\,\alpha_n^I\,\Omega^{-1}\ (n\neq 0)\,.$$

(c) Show that $\Omega\, X^-(\tau,\sigma)\,\Omega^{-1} = X^-(\tau,\pi-\sigma)$. Since $\Omega\, X^+(\tau,\sigma)\,\Omega^{-1} = X^+(\tau,\ \pi-\sigma)$, equation (1) actually holds for all string coordinates. We say that orientation reversal is a symmetry of open string theory because it leaves the open string Hamiltonian H invariant: $\Omega H\Omega^{-1} = H$. Explain why this is true.

(d) Assume that the ground states are twist invariant:

$$\Omega\,|p^+,\vec{p}_T\rangle = \Omega^{-1}\,|p^+,\vec{p}_T\rangle = |p^+,\vec{p}_T\rangle\,.$$

List the open string states for $N^\perp\leq 3$, and give their twist eigenvalues. Prove that, in general,

$$\Omega = (-1)^{N^\perp}\,.$$

(e) A state is said to be *unoriented* if it is invariant under twist. If you are commissioned to build a theory of unoriented open strings, which of the states in part (d) would you have to discard? In general, which levels of the original string state space must be discarded?

Problem 12.13 Tachyon potentials.

Consider scalar field theories of the form

$$S = \int d^Dx\Big(-\frac{1}{2}\partial_\mu\phi\partial^\mu\phi - V(\phi)\Big)\,. \tag{1}$$

We will examine three different scalar potentials:

$$V_1(\phi) = \frac{1}{\alpha'}\,\frac{1}{3\phi_0}\,(\phi-\phi_0)^2\Big(\phi+\frac{1}{2}\phi_0\Big), \tag{2a}$$

$$V_2(\phi) = -\frac{1}{4\alpha'}\,\phi^2\,\ln\Big(\frac{\phi^2}{\phi_0^2}\Big), \tag{2b}$$

$$V_3(\phi) = \frac{1}{8\alpha'}\,\phi_0^2\,\Big(\frac{\phi^2}{\phi_0^2}-1\Big)^2, \tag{2c}$$

where ϕ_0 is a (positive) constant. For each of the three potentials V_i, do the following.

(a) Plot $V_i(\phi)$ as a function of ϕ.
(b) Find the critical points of the potential and the values of the potentials at those points. Each critical point represents a possible background for the scalar field theory.
(c) At each critical point $\bar{\phi}$ expand the action for fluctuations of ϕ around this point, that is, let $\phi = \bar{\phi} + \psi$ where the fluctuation ψ is small. The quadratic term in ψ (with no derivatives) can be used to read the mass of the scalar particle. Give the mass of the scalar particle for each critical point.

The potential V_1 is a rough model for the tachyon potential on an unstable D-brane. V_2 is an exact (effective) tachyon potential on an unstable D-brane. The potential V_3 is a rough model for the superstring tachyon potential on the world-volume of a D-brane and a coincident anti-D-brane.

13 Relativistic quantum closed strings

Except for shared position and momentum zero modes, the operator content of quantum closed strings can be viewed as two commuting copies of the open string operators. Even in the light-cone gauge, the reparameterization invariance cannot be fully fixed: there is no natural way to choose a starting point for a closed string. As a result, the closed string spectrum is subject to the constraint $L_0^\perp - \tilde{L}_0^\perp = 0$, which selects the states that are invariant under rigid rotations of the string. We find that the massless closed string quantum states include one-particle graviton states, making string theory a quantum gravity theory. Additionally, we find massless Kalb–Ramond and dilaton states. The dilaton state controls the strength of string interactions. We study closed strings on the orbifold $\mathbb{R}^1/\mathbb{Z}_2$.

13.1 Mode expansions and commutation relations

When it was first discovered, string theory was thought to be a theory of strongly interacting particles – a theory of hadrons. The consistency of open string theory required the inclusion of closed strings. But there was a problem with closed strings: among the excitations of closed strings there were massless states with spin two. No known hadron had these properties. Despite much effort, all attempts to eliminate these closed string states from the spectrum failed.

It turns out that these massless states can be identified as graviton states, and physicists eventually realized that closed string theory could be a theory of quantum gravity. In this chapter we quantize the relativistic closed string and see how graviton states emerge. Much of the quantization procedure will resemble our quantization of the open string in Chapter 12, but there are a number of new features. Let us begin by recalling some of the important facts about closed strings that we learned in Chapter 9. We considered at that time a family of gauges (see (9.27)) defined by the conditions

$$n \cdot X = \alpha'(n \cdot p)\,\tau\,, \quad n \cdot p = 2\pi\, n \cdot \mathcal{P}^\tau\,. \tag{13.1}$$

The second condition implies that the parameter σ spans an interval of length 2π:

$$\sigma \in [0, 2\pi]\,. \tag{13.2}$$

Here, $\sigma = 0$ and $\sigma = 2\pi$ represent the *same* point on the closed string. The range $\sigma \in [0, 2\pi]$ for closed strings is twice the range $\sigma \in [0, \pi]$ used for open strings. We found

that the conditions (13.1) do not fully fix the parameterization of the closed strings. Unlike open strings, closed strings do not have a special point that can be selected as $\sigma = 0$. We used this arbitrariness to our advantage: once we selected some $\sigma = 0$ on one closed string, we could impose the constraint $X' \cdot \dot{X} = 0$ by suitably choosing the $\sigma = 0$ point on all of the other closed strings on the world-sheet. After this, we still had the ability to let $\sigma \to \sigma + \sigma_0$, with some constant σ_0 that is the same for all strings. This rigid rotation of the lines of constant σ is a reparameterization invariance of the closed string action that cannot be fixed. When we build the quantum states of the closed string, this will result in a constraint on states.

The condition $X' \cdot \dot{X} = 0$, together with the parameterization conditions (13.1), implied $X'^2 + \dot{X}^2 = 0$. We thus obtained the familiar conditions

$$(\dot{X} \pm X')^2 = 0\,, \tag{13.3}$$

and the momentum densities became simple derivatives of the coordinates:

$$\mathcal{P}^{\sigma\mu} = -\frac{1}{2\pi\alpha'} X'^{\mu}, \quad \mathcal{P}^{\tau\mu} = \frac{1}{2\pi\alpha'} \dot{X}^{\mu}. \tag{13.4}$$

Finally, all the string coordinates were seen to satisfy the wave equation:

$$\Big(\frac{\partial^2}{\partial\tau^2} - \frac{\partial^2}{\partial\sigma^2}\Big) X^{\mu} = 0\,. \tag{13.5}$$

Let us now consider the classical solution to the equation of motion for the closed string. The general solution to the wave equation is

$$X^{\mu}(\tau,\sigma) = X^{\mu}_{L}(\tau+\sigma) + X^{\mu}_{R}(\tau-\sigma)\,, \tag{13.6}$$

where X^{μ}_{L} (the L stands for left-moving) is a wave moving towards more negative σ and X^{μ}_{R} (the R stands for right-moving) is a wave moving towards more positive σ. For open strings, the left-moving and right-moving waves were related to each other by the boundary conditions at the endpoints. The closed string has no endpoints, but we have a periodicity condition to work with. The parameter space (τ,σ) for closed strings is a cylinder, so, to describe closed strings properly we compactify the world-sheet coordinate σ:

$$\sigma \sim \sigma + 2\pi\,. \tag{13.7}$$

Two points on the world-sheet whose difference of σ coordinates is a multiple of 2π are the same point. We can in fact use any interval of the form $[\sigma_0, \sigma_0 + 2\pi]$ to describe the closed strings; the choice in (13.2) is one of the possible choices. When we include the τ coordinate, the identification of points on the parameter space is given by

$$(\tau,\sigma) \sim (\tau,\sigma+2\pi)\,. \tag{13.8}$$

We demand that X^{μ} assumes the same value at any two coordinates that represent the same point on the parameter space:

$$X^\mu(\tau,\sigma) = X^\mu(\tau,\sigma+2\pi)\,, \quad \text{for all } \tau \text{ and } \sigma\,. \tag{13.9}$$

This condition is both easier to deal with and easier to interpret than the naive condition $X^\mu(\tau,0) = X^\mu(\tau,2\pi)$. The periodicity condition (13.9) is appropriate for strings that propagate in a simply connected space, a space in which every closed string can be continuously shrunk to a point. Minkowski space, for example, is simply connected. If a spatial direction is curled up into a circle, closed strings wrapped around the circle cannot be shrunk away. Space is then not simply connected, and the coordinate along the circle is not single valued. For such a coordinate, the periodicity condition (13.9) must be modified (see Chapter 17).

We will now show that the periodicity condition (13.9) induces a small but significant constraint that relates the left-moving and the right-moving waves. Let us define two new variables,

$$\begin{aligned} u &= \tau + \sigma\,, \\ v &= \tau - \sigma\,. \end{aligned} \tag{13.10}$$

In terms of these variables, equation (13.6) becomes

$$X^\mu = X^\mu_L(u) + X^\mu_R(v)\,. \tag{13.11}$$

When $\sigma \to \sigma + 2\pi$, the variables u and v increase and decrease by 2π, respectively. As a result, the periodicity condition (13.9) gives

$$X^\mu_L(u) + X^\mu_R(v) = X^\mu_L(u+2\pi) + X^\mu_R(v-2\pi)\,, \tag{13.12}$$

or, equivalently,

$$X^\mu_L(u+2\pi) - X^\mu_L(u) = X^\mu_R(v) - X^\mu_R(v-2\pi)\,. \tag{13.13}$$

This equation establishes that the left-moving and right-moving waves are in fact dependent on each other: if one fails to be periodic, the other has to fail by the same amount. Since u and v are independent variables, both the u derivative of the right-hand side and the v derivative of the left-hand side must vanish. As a consequence, we find that both $X^{\mu\prime}_L(u)$ and $X^{\mu\prime}_R(v)$ are strictly periodic functions with period 2π (for functions of a single variable primes denote derivatives with respect to the argument). We can therefore write the mode expansions

$$\begin{aligned} X^{\mu\prime}_L(u) &= \sqrt{\frac{\alpha'}{2}} \sum_{n\in\mathbb{Z}} \bar{\alpha}^\mu_n e^{-inu}, \\ X^{\mu\prime}_R(v) &= \sqrt{\frac{\alpha'}{2}} \sum_{n\in\mathbb{Z}} \alpha^\mu_n e^{-inv}\,. \end{aligned} \tag{13.14}$$

A set of barred α modes was introduced for the expansion of $X^{\mu\prime}_L(u)$. Even though they are written identically, the unbarred α modes used in the expansion of $X^{\mu\prime}_R(v)$ have no

relation to the open string modes of Chapter 12. In closed string theory we need two sets of α modes, barred and unbarred. We integrate equations (13.14) to find

$$X_L^\mu(u) = \frac{1}{2}\, x_0^{L\mu} + \sqrt{\frac{\alpha'}{2}}\, \bar{\alpha}_0^\mu\, u + i\sqrt{\frac{\alpha'}{2}} \sum_{n\neq 0} \frac{\bar{\alpha}_n^\mu}{n} e^{-inu},$$
$$X_R^\mu(v) = \frac{1}{2}\, x_0^{R\mu} + \sqrt{\frac{\alpha'}{2}}\, \alpha_0^\mu\, v + i\sqrt{\frac{\alpha'}{2}} \sum_{n\neq 0} \frac{\alpha_n^\mu}{n} e^{-inv}, \tag{13.15}$$

where the coordinate zero modes $x_0^{L\mu}$ and $x_0^{R\mu}$ have appeared as constants of integration. These are somewhat puzzling, after all, in open string theory there was a single coordinate zero mode whose canonical conjugate was the momentum of the string. We will see that only the sum of the two zero modes plays a role here. If the space is not simply connected, however, each of the coordinate zero modes plays a role, as we will see in Chapter 17.

The aperiodicity of X_R^μ and of X_L^μ is a consequence of the linear terms appearing in (13.15). Condition (13.13) constrains these terms giving

$$2\pi \sqrt{\frac{\alpha'}{2}}\, \bar{\alpha}_0^\mu = 2\pi \sqrt{\frac{\alpha'}{2}}\, \alpha_0^\mu\,, \tag{13.16}$$

and therefore

$$\bar{\alpha}_0^\mu = \alpha_0^\mu\,. \tag{13.17}$$

Owing to this equality, quantum closed string theory has only *one* momentum operator. As we will soon see, this means that canonical quantization works consistently with only *one* coordinate zero mode operator.

We can now assemble the mode expansion for $X^\mu(\tau,\sigma)$ by substituting (13.15) into (13.6):

$$X^\mu(\tau,\sigma) = \frac{1}{2}\, x_0^{L\mu} + \sqrt{\frac{\alpha'}{2}}\, \bar{\alpha}_0^\mu(\tau+\sigma) + i\sqrt{\frac{\alpha'}{2}} \sum_{n\neq 0} \frac{\bar{\alpha}_n^\mu}{n} e^{-in(\tau+\sigma)}$$
$$+ \frac{1}{2}\, x_0^{R\mu} + \sqrt{\frac{\alpha'}{2}}\, \alpha_0^\mu(\tau-\sigma) + i\sqrt{\frac{\alpha'}{2}} \sum_{n\neq 0} \frac{\alpha_n^\mu}{n} e^{-in(\tau-\sigma)}\,. \tag{13.18}$$

With $\bar{\alpha}_0^\mu = \alpha_0^\mu$, we find

$$X^\mu(\tau,\sigma) = \frac{1}{2}(x_0^{L\mu} + x_0^{R\mu}) + \sqrt{2\alpha'}\, \alpha_0^\mu\, \tau + i\sqrt{\frac{\alpha'}{2}} \sum_{n\neq 0} \frac{e^{-in\tau}}{n} (\alpha_n^\mu e^{in\sigma} + \bar{\alpha}_n^\mu e^{-in\sigma})\,. \tag{13.19}$$

As expected, X^μ is a periodic function of σ with period 2π. The canonically conjugate momentum density is

$$\mathcal{P}^{\tau\mu}(\tau,\sigma) = \frac{1}{2\pi\alpha'} \dot{X}^\mu(\tau,\sigma) = \frac{1}{2\pi\alpha'} (\sqrt{2\alpha'}\, \alpha_0^\mu + \cdots)\,, \tag{13.20}$$

where the dots represent the terms in $\dot{X}^\mu$ that integrate to zero over the interval $\sigma \in [0, 2\pi]$ and therefore do not contribute to the evaluation of the total momentum:

$$p^\mu = \int_0^{2\pi} \mathcal{P}^{\tau\mu}(\tau,\sigma)d\sigma = \frac{1}{2\pi\alpha'}\int_0^{2\pi} d\sigma\,\sqrt{2\alpha'}\,\alpha_0^\mu = \sqrt{\frac{2}{\alpha'}}\alpha_0^\mu\,. \tag{13.21}$$

Thus we have the relation

$$\alpha_0^\mu = \sqrt{\frac{\alpha'}{2}}\;p^\mu\,. \tag{13.22}$$

This differs from the corresponding open string result (12.46) by a factor of two, but the idea is the same: α_0^μ is proportional to the spacetime momentum carried by the string.

There is only one momentum variable, and thus in the quantum theory there is only one momentum operator. We should also have only one conjugate coordinate zero mode. Thus, despite our left–right decomposition of the solution to the wave equation, x_0^L and x_0^R cannot both be independent variables. Only their sum appears in (13.19), so it must be the sum that is the relevant coordinate zero mode. Without loss of generality, we set

$$x_0^{L\mu} = x_0^{R\mu} \equiv x_0^\mu\,, \tag{13.23}$$

and equation (13.19) takes its final form:

$$X^\mu(\tau,\sigma) = x_0^\mu + \sqrt{2\alpha'}\,\alpha_0^\mu\tau + i\sqrt{\frac{\alpha'}{2}}\sum_{n\neq 0}\frac{e^{-in\tau}}{n}(\alpha_n^\mu e^{in\sigma} + \bar{\alpha}_n^\mu e^{-in\sigma})\,. \tag{13.24}$$

It is convenient at this stage to record the τ and σ derivatives of the coordinates. With the help of (13.6) we note that

$$\begin{aligned}\dot{X}^\mu &= X_L^{\mu\prime}(\tau+\sigma) + X_R^{\mu\prime}(\tau-\sigma)\,,\\ X^{\mu\prime} &= X_L^{\mu\prime}(\tau+\sigma) - X_R^{\mu\prime}(\tau-\sigma)\,.\end{aligned} \tag{13.25}$$

Adding and subtracting these equations, and using (13.14), we find

$$\begin{aligned}\dot{X}^\mu + X^{\mu\prime} &= 2X_L^{\mu\prime}(\tau+\sigma) = \sqrt{2\alpha'}\sum_{n\in\mathbb{Z}}\bar{\alpha}_n^\mu\, e^{-in(\tau+\sigma)}\,,\\ \dot{X}^\mu - X^{\mu\prime} &= 2X_R^{\mu\prime}(\tau-\sigma) = \sqrt{2\alpha'}\sum_{n\in\mathbb{Z}}\alpha_n^\mu e^{-in(\tau-\sigma)}\,.\end{aligned} \tag{13.26}$$

Note that the barred oscillators do not mix with the unbarred oscillators in these combinations of derivatives. We have tailored the normalization constants to arrive at the above relations. They are completely analogous to the open string expansions (12.33). This will allow us to obtain some closed string commutators without doing any new computations.

Let us now turn to the quantization of the closed string theory. The canonical commutation relations take the same form as in open string theory. For the transverse light-cone coordinates and momenta we set

$$\left[X^I(\tau,\sigma)\,,\mathcal{P}^{\tau J}(\tau,\sigma')\right] = i\delta(\sigma-\sigma')\eta^{IJ}\,, \tag{13.27}$$

and, as usual, we set to zero the commutator of coordinates with coordinates and the commutator of momenta with momenta. For zero modes, we also have $[x_0^-\,,\,p^+] = -i$. Since the commutation relations did not change, equations (12.30) and (12.31) are valid here. The first of these is

$$[\,(\dot{X}^I \pm X^{I\prime})(\tau,\sigma)\,,\,(\dot{X}^J \pm X^{J\prime})(\tau,\sigma')\,] = \pm 4\pi\alpha' i\eta^{IJ}\frac{d}{d\sigma}\delta(\sigma-\sigma')\,. \tag{13.28}$$

In fact, this time the situation is simpler. Equations (13.28) hold for $\sigma,\sigma' \in [0,2\pi]$ since the string coordinates are defined for this full interval. Moreover, the mode expansions (13.26) also hold for the full interval. Since the combinations of derivatives take the same exact form as they did for open strings, the above equation leads to identical looking commutation relations. The oscillators, however, are barred when we use the top sign and unbarred when we use the lower sign. The result is therefore

$$\begin{aligned}\left[\bar{\alpha}_m^I,\bar{\alpha}_n^J\right] &= m\,\delta_{m+n,0}\,\eta^{IJ}\,,\\ \left[\alpha_m^I,\alpha_n^J\right] &= m\,\delta_{m+n,0}\,\eta^{IJ}\,.\end{aligned} \tag{13.29}$$

On account of the expansions (13.26), we call the $\bar{\alpha}$ operators left-moving operators, and we call the α operators right-moving operators. Each of these sets matches the operator content of an open string theory. The commutation relations also take the form we would expect from open string theory. Closed string theory thus has the operator content of two copies of open string theory, except for zero modes. The momentum zero modes are equal ($\alpha_0^I = \bar{\alpha}_0^I$), and there is only one set of coordinate zero modes x_0^I, x_0^-.

Equation (12.31) states that combinations of derivatives with opposite signs commute. In the present case this leads to the result that left-moving and right-moving oscillators commute:

$$\left[\alpha_m^I\,,\bar{\alpha}_n^J\right] = 0\,. \tag{13.30}$$

We can define canonical creation and annihilation operators just as we did for open strings:

$$\begin{aligned}\alpha_n^I &= a_n^I\sqrt{n} \quad\text{and}\quad \alpha_{-n}^I = a_n^{I\dagger}\sqrt{n}\,, \quad n\geq 1\,,\\ \bar{\alpha}_n^I &= \bar{a}_n^I\sqrt{n} \quad\text{and}\quad \bar{\alpha}_{-n}^I = \bar{a}_n^{I\dagger}\sqrt{n}\,, \quad n\geq 1\,.\end{aligned} \tag{13.31}$$

The nonvanishing commutation relations are then the expected ones:

$$\left[\,\bar{a}_m^I\,,\,\bar{a}_n^{J\dagger}\,\right] = \delta_{m,n}\,\eta^{IJ}\,, \qquad \left[\,a_m^I\,,\,a_n^{J\dagger}\,\right] = \delta_{m,n}\,\eta^{IJ}\,. \tag{13.32}$$

The commutators that involve x_0^I can be found following steps analogous to those used for open strings. This time (Problem 13.1) we find that $[x_0^I, \alpha_n^J]$ and $[x_0^I, \bar{\alpha}_n^J]$ vanish when $n \neq 0$, and

$$\left[\,x_0^I\,,\alpha_0^J\,\right] = \left[\,x_0^I\,,\bar{\alpha}_0^J\,\right] = i\sqrt{\frac{\alpha'}{2}}\;\eta^{IJ} \;\longrightarrow\; \left[\,x_0^I\,,\;p^J\,\right] = i\eta^{IJ}\,, \tag{13.33}$$

where the expression to the right arises because of (13.22).

What is the light-cone closed string Hamiltonian? We know that p^- generates X^+ translations, and that, in addition, $X^+ = \alpha' p^+ \tau$. As a result, $\partial_\tau = \alpha' p^+ \partial_{X^+}$, and the Hamiltonian must be given by

$$H = \alpha' p^+ p^-\,. \tag{13.34}$$

In order to find the normal-ordered version of this Hamiltonian, we now turn to the transverse Virasoro operators of closed string theory.

13.2 Closed string Virasoro operators

We learned in Chapter 12 that the open string transverse Virasoro operators are essentially the modes α_n^- of the light-cone coordinate X^-. For closed string coordinates we have two types of modes, barred and unbarred. This is also true for the closed string X^- coordinates: we have α_n^- and $\bar{\alpha}_n^-$ modes, and therefore we expect to have two sets of Virasoro operators. On account of (13.17), however, we have $\alpha_0^- = \bar{\alpha}_0^-$, so a surprise awaits us with regards to the Virasoro operators with mode number zero.

To begin our analysis we need an expression that relates X^- to the transverse coordinates. The requisite formula is (9.65), with $\beta = 1$, as appropriate for closed strings:

$$\dot{X}^- \pm X^{-\prime} = \frac{1}{\alpha'}\frac{1}{2p^+}(\dot{X}^I \pm X^{I\prime})^2\,. \tag{13.35}$$

We define Virasoro operators following the pattern in equation (9.79):

$$\begin{aligned}(\dot{X}^I + X^{I\prime})^2 &= 4\alpha' \sum_{n\in\mathbb{Z}}\Big(\frac{1}{2}\sum_{p\in\mathbb{Z}}\bar{\alpha}_p^I\bar{\alpha}_{n-p}^I\Big)e^{-in(\tau+\sigma)} \equiv 4\alpha'\sum_{n\in\mathbb{Z}}\bar{L}_n^{\perp}e^{-in(\tau+\sigma)}\,,\\ (\dot{X}^I - X^{I\prime})^2 &= 4\alpha' \sum_{n\in\mathbb{Z}}\Big(\frac{1}{2}\sum_{p\in\mathbb{Z}}\alpha_p^I\alpha_{n-p}^I\Big)e^{-in(\tau-\sigma)} \equiv 4\alpha'\sum_{n\in\mathbb{Z}}L_n^{\perp}e^{-in(\tau-\sigma)}\,.\end{aligned} \tag{13.36}$$

In each of the above lines, the equality requires a small calculation using (13.26), and the second relation is a definition. More explicitly,

$$\bar{L}_n^\perp = \frac{1}{2}\sum_{p\in\mathbb{Z}} \bar{\alpha}_p^I \bar{\alpha}_{n-p}^I\,, \quad L_n^\perp = \frac{1}{2}\sum_{p\in\mathbb{Z}} \alpha_p^I \alpha_{n-p}^I\,. \tag{13.37}$$

These are the two sets of Virasoro operators of closed string theory. Plugging the definitions in (13.36) back into (13.35) we obtain

$$\dot{X}^- + X^{-\prime} = \frac{2}{p^+}\sum_{n\in\mathbb{Z}} \bar{L}_n^\perp\, e^{-in(\tau+\sigma)}\,, \quad \dot{X}^- - X^{-\prime} = \frac{2}{p^+}\sum_{n\in\mathbb{Z}} L_n^\perp\, e^{-in(\tau-\sigma)}\,. \tag{13.38}$$

On the other hand, the derivatives of X^-, as those of any other closed string coordinate, can be expanded along the lines of (13.26) to give

$$\dot{X}^- + X^{-\prime} = \sqrt{2\alpha'}\sum_{n\in\mathbb{Z}} \bar{\alpha}_n^-\, e^{-in(\tau+\sigma)}\,, \quad \dot{X}^- - X^{-\prime} = \sqrt{2\alpha'}\sum_{n\in\mathbb{Z}} \alpha_n^-\, e^{-in(\tau-\sigma)}\,. \tag{13.39}$$

We compare equations (13.38) and (13.39) to read the expressions for the minus oscillators:

$$\sqrt{2\alpha'}\,\bar{\alpha}_n^- = \frac{2}{p^+}\,\bar{L}_n^\perp\,, \quad \sqrt{2\alpha'}\,\alpha_n^- = \frac{2}{p^+}\,L_n^\perp\,. \tag{13.40}$$

For $n = 0$, however, there is a constraint. Since $\alpha_0^- = \bar{\alpha}_0^-$, we have the *level-matching* condition

$$L_0^\perp = \bar{L}_0^\perp\,. \tag{13.41}$$

If you look at the definitions of $\bar{L}_0^\perp$ and $L_0^\perp$ in (13.37), you will realize that these two operators are clearly very different from each other. What does it mean that they must be equal? Since operators are ultimately defined by their action on states, the meaning of the equality (13.41) is that any state $|\lambda, \bar{\lambda}\rangle$ of the closed string must satisfy $L_0^\perp|\lambda, \bar{\lambda}\rangle = \bar{L}_0^\perp|\lambda, \bar{\lambda}\rangle$. This is therefore a constraint on the state space of the theory: "states" that do not satisfy this constraint do not in fact belong to the state space.

To fix the ordering ambiguities in the operators $\bar{L}_0^\perp$ and $L_0^\perp$ we define them to be ordered operators without any additional constants:

$$\bar{L}_0^\perp = \frac{\alpha'}{4}p^I p^I + \bar{N}^\perp\,, \quad L_0^\perp = \frac{\alpha'}{4}p^I p^I + N^\perp\,. \tag{13.42}$$

Here $\bar{N}^\perp$ and $N^\perp$ are the number operators that are associated with the barred and unbarred operators, respectively:

$$\bar{N}^\perp \equiv \sum_{n=1}^{\infty} n\bar{a}_n^{I\dagger}\,\bar{a}_n^I\,, \quad N^\perp \equiv \sum_{n=1}^{\infty} n a_n^{I\dagger}\,a_n^I\,. \tag{13.43}$$

While we will not go through the trouble of proving it, the critical dimension for closed strings turns out to be $D = 26$. This follows from the requirement that the quantum theory

be Lorentz invariant. It is no coincidence that the critical dimension for closed strings coincides with the critical dimension for open strings. It means that both types of strings can coexist. In fact, since open strings can, in general, close to form closed strings, it would have been quite strange if the critical dimensions did not agree.

The constant ambiguities due to the ordering of $\bar{L}_0^\perp$ and $L_0^\perp$ are also fixed by the condition of Lorentz invariance, just as it happened for open strings. The answer could be anticipated, since the left and right sectors of closed string theory behave like open strings. In addition, the naive argument based on zeta functions suggests that the ordering constants for $L_0^\perp$ and $\bar{L}_0^\perp$ are the same and equal to that for the $L_0^\perp$ operator of the open string. These constants are included in the relation between α_0^- and $L_0^\perp$ and in the corresponding barred relation. Therefore equations (13.40), for $n = 0$, become

$$\sqrt{2\alpha'}\,\bar{\alpha}_0^- = \frac{2}{p^+}(\bar{L}_0^\perp - 1)\,, \quad \sqrt{2\alpha'}\,\alpha_0^- = \frac{2}{p^+}(L_0^\perp - 1)\,. \tag{13.44}$$

The level-matching constraint $L_0^\perp = \bar{L}_0^\perp$, which emerged from $\alpha_0^- = \bar{\alpha}_0^-$, remains unchanged by the constant shifts. Using (13.42), this constraint can be written more simply as

$$N^\perp = \bar{N}^\perp\,. \tag{13.45}$$

Averaging the two expressions for α_0^- in (13.44), we can find a symmetric expression:

$$\sqrt{2\alpha'}\alpha_0^- \equiv \frac{1}{p^+}(L_0^\perp + \bar{L}_0^\perp - 2) = \alpha' p^-\,, \tag{13.46}$$

where the relation to p^- follows from (13.22). With p^- known, we can calculate the mass-squared:

$$M^2 = -p^2 = 2p^+p^- - p^I p^I = \frac{2}{\alpha'}(L_0^\perp + \bar{L}_0^\perp - 2) - p^I p^I\,. \tag{13.47}$$

Substituting the values of $\bar{L}_0^\perp$ and $L_0^\perp$ given in (13.42) yields

$$M^2 = \frac{2}{\alpha'}\left(N^\perp + \bar{N}^\perp - 2\right). \tag{13.48}$$

This is the mass formula for closed string states. The closed string Hamiltonian (13.34) can be written in terms of Virasoro operators using (13.46). The result is very simple:

$$H = \alpha' p^+ p^- = L_0^\perp + \bar{L}_0^\perp - 2\,. \tag{13.49}$$

This Hamiltonian is the sum of an "open string" Hamiltonian $L_0^\perp - 1$ for the right-moving operators, and an "open string" Hamiltonian $\bar{L}_0^\perp - 1$ for the left-moving operators. Using (13.42) the Hamiltonian can be written as

$$H = \frac{\alpha'}{2} p^I p^I + N^\perp + \bar{N}^\perp - 2\,. \tag{13.50}$$

Both the $L_m^\perp$ operators and the $\bar{L}_m^\perp$ operators satisfy the Virasoro algebra (12.133). In addition, the commutators between barred and unbarred Virasoro operators vanish. So, the full set of closed string Virasoro operators define two commuting Virasoro algebras.

We conclude this section with a study of the Virasoro action on closed string coordinates. The commutation of the closed string Virasoro operators with the oscillators follows the pattern of equation (12.118). We have

$$\left[\, \bar{L}_m^\perp \,,\, \bar{\alpha}_n^J \,\right] = -n\, \bar{\alpha}_{m+n}^J\,, \qquad \left[\, L_m^\perp \,,\, \alpha_n^J \,\right] = -n\, \alpha_{m+n}^J\,, \tag{13.51}$$

and, in addition,

$$\left[\, L_m^\perp \,,\, \bar{\alpha}_n^J \,\right] = \left[\, \bar{L}_m^\perp \,,\, \alpha_n^J \,\right] = 0\,. \tag{13.52}$$

On the other hand, both $\bar{L}_m^\perp$ and $L_m^\perp$ have a nontrivial commutator with x_0^I:

Quick calculation 13.1 Verify that

$$\left[\, \bar{L}_m^\perp \,,\, x_0^I \,\right] = -i \sqrt{\frac{\alpha'}{2}}\, \bar{\alpha}_m^I\,, \qquad \left[\, L_m^\perp \,,\, x_0^I \,\right] = -i \sqrt{\frac{\alpha'}{2}}\, \alpha_m^I\,. \tag{13.53}$$

Let us focus here only on the action of $L_0^\perp$ and $\bar{L}_0^\perp$ on the string coordinates. The required formulae are obtained in the following exercise:

Quick calculation 13.2 Verify that

$$\left[\, \bar{L}_0^\perp \,,\, X^I(\tau,\sigma) \,\right] = -\frac{i}{2}\,(\dot{X}^I + X^{I\prime})\,, \qquad \left[\, L_0^\perp \,,\, X^I(\tau,\sigma) \,\right] = -\frac{i}{2}\,(\dot{X}^I - X^{I\prime})\,. \tag{13.54}$$

Adding the two equations in (13.54), we find

$$\left[\, L_0^\perp + \bar{L}_0^\perp \,,\; X^I(\tau,\sigma) \,\right] = -i\, \frac{\partial X^I}{\partial \tau}\,. \tag{13.55}$$

This equation is consistent with the Heisenberg equation of motion for X^I, since the closed string Hamiltonian is $(L_0^\perp + \bar{L}_0^\perp - 2)$. Subtracting the two equations in (13.54), we find a more surprising result:

$$\left[\, L_0^\perp - \bar{L}_0^\perp \,,\, X^I(\tau,\sigma) \,\right] = i\, \frac{\partial X^I}{\partial \sigma}\,. \tag{13.56}$$

This equation shows that $L_0^\perp - \bar{L}_0^\perp$ generates constant translations along the string. Indeed, for infinitesimal ϵ,

$$X^I(\tau,\sigma) + \left[\, -i\epsilon (L_0^\perp - \bar{L}_0^\perp)\,,\, X^I(\tau,\sigma) \,\right] = X^I(\tau,\sigma+\epsilon)\,. \tag{13.57}$$

More generally, a finite translation along the string can be obtained by acting on the string coordinate with exponentials of $L_0^\perp - \bar{L}_0^\perp$. Writing

$$P \equiv L_0^\perp - \bar{L}_0^\perp\,, \tag{13.58}$$

we find that (Problem 13.3)

$$e^{-iP\sigma_0}\, X^I(\tau,\sigma)\, e^{iP\sigma_0} = X^I(\tau,\sigma+\sigma_0)\,, \tag{13.59}$$

for any finite σ_0. For $\sigma_0 = \epsilon$, infinitesimal, this general result reduces to (13.57). The operator P generates the one reparameterization symmetry that cannot be fixed even in the light-cone gauge. Since P annihilates all closed string states (see (13.41)), we conclude that closed string states are invariant under rigid σ translations. By this we mean that $\exp(-iP\sigma_0)|\Psi\rangle = |\Psi\rangle$, for any closed string state $|\Psi\rangle$.

In the two-dimensional (τ,σ) parameter space of the closed string world-sheet, the operator $L_0^\perp + \bar{L}_0^\perp$ is a generator of τ translations. It is therefore a world-sheet energy. Since the gauge condition relates τ to the light-cone time, this world-sheet energy turns out to give us the spacetime Hamiltonian, the generator of light-cone time evolution. The other combination $L_0^\perp - \bar{L}_0^\perp = P$ generates translations along the world-sheet coordinate σ. It can therefore be viewed as a *world-sheet* momentum. This momentum should not be confused with the spacetime momentum of the string. For closed string states the world-sheet momentum must in fact vanish, and this is a nontrivial constraint. States with nonzero momentum along σ can be built, but they do not belong to the closed string state space.

13.3 Closed string state space

We are now ready to build the state space of the quantum closed string. The ground states are $|p^+, \vec{p}_T\rangle$ and they are annihilated by both the left-moving and the right-moving annihilation operators. To generate all of the basis states we must act on the ground states with the creation operators $a_n^{I\dagger}$ and $\bar{a}_n^{I\dagger}$. The general *candidate* basis vector is

$$|\lambda,\bar{\lambda}\rangle = \Big[\prod_{n=1}^{\infty}\prod_{I=2}^{25}(a_n^{I\dagger})^{\lambda_{n,I}}\Big]\times\Big[\prod_{m=1}^{\infty}\prod_{J=2}^{25}(\bar{a}_m^{J\dagger})^{\bar{\lambda}_{m,J}}\Big]|p^+,\vec{p}_T\rangle. \tag{13.60}$$

Just as with open strings, the occupation numbers $\lambda_{n,I}$ and $\bar{\lambda}_{n,I}$ are non-negative integers. The number operators act on $|\lambda,\bar{\lambda}\rangle$ with eigenvalues

$$N^\perp = \sum_{n=1}^{\infty}\sum_{I=2}^{25} n\lambda_{n,I}\,, \qquad \bar{N}^\perp = \sum_{m=1}^{\infty}\sum_{J=2}^{25} m\bar{\lambda}_{m,J}\,. \tag{13.61}$$

Except for the momentum labels, the above states are those that one would obtain by combining *multiplicatively* arbitrary states built from the left-moving and from the right-moving operators (compare with (12.162)). Not all of the states in (13.60) belong to the closed string state space. The constraint (13.45) must be satisfied by the true states of the theory. A basis vector $|\lambda,\bar{\lambda}\rangle$ belongs to the state space *if and only if* it satisfies the level-matching constraint

$$N^\perp = \bar{N}^\perp\,. \tag{13.62}$$

Table 13.1 The states with $N^\perp + \bar{N}^\perp \leq 2$ in the closed string spectrum

$N^\perp, \bar{N}^\perp$	$\lvert\lambda, \bar{\lambda}\rangle$	$\frac{1}{2}\alpha' M^2$	Number of states	Wavefunction
$0, 0$	$\lvert p^+, \vec{p}_T\rangle$	-2	1	$\psi(\tau, p^+, \vec{p}_T)$
$1, 1$	$a_1^{I\dagger}\bar{a}_1^{J\dagger}\lvert p^+, \vec{p}_T\rangle$	0	$(D-2)^2$	$\psi_{IJ}(\tau, p^+, \vec{p}_T)$

This constraint cannot be "solved". It is clear that eliminating some oscillator from the list of operators that can act on the ground states is no help whatsoever. A better strategy would be to select a particular oscillator and attempt to use it, as many times as needed, to fix any state that does not satisfy (13.62). Not even this is possible, because any oscillator can only contribute a positive amount to the number operator. The constraint (13.62) must be implemented in a case-by-case fashion. The masses of the states are obtained from (13.48):

$$\frac{1}{2}\alpha' M^2 = N^\perp + \bar{N}^\perp - 2\,. \tag{13.63}$$

Let us identify the first few basis states, give their masses, and explain what fields they represent. The results are tabulated in Table 13.1.

The ground states in the first row of Table 13.1 are the one-particle states of a quantum scalar field. For such states $N^\perp = \bar{N}^\perp = 0$ and $M^2 = -4/\alpha' < 0$, so these are closed string tachyons; they are in fact completely analogous to the tachyons of open string theory. The mass-squared of the closed string tachyon is four times larger than that of the open string tachyon. The closed string tachyon is far less understood than the open string tachyon. In particular, the closed string tachyon potential has not yet been calculated reliably. The instabilities associated with closed string tachyons are expected to be instabilities of spacetime itself. They remain largely mysterious.

The next excited states must be built with *two* oscillators acting on the ground states. This is because we must satisfy the constraint $N^\perp = \bar{N}^\perp$. One oscillator must be from the left-sector and one from the right-sector, both with the lowest possible mode number – mode number one. This gives the states described in the second line of the table. All these states have $M^2 = 0$, so they are of great interest. Since I and J are completely arbitrary labels attached to *different* oscillators, the number of states is $(D-2)^2$.

Let us consider the general state of fixed momentum at the massless level. We write it as

$$\sum_{I,J} R_{IJ}\; a_1^{I\dagger}\bar{a}_1^{J\dagger}\lvert p^+, \vec{p}_T\rangle\,. \tag{13.64}$$

Here R_{IJ} are the elements of an arbitrary square matrix of size $(D-2)$. Any square matrix can be decomposed into its symmetric part and its antisymmetric part:

$$R_{IJ} = \frac{1}{2}\,(R_{IJ} + R_{JI}) + \frac{1}{2}\,(R_{IJ} - R_{JI}) \equiv S_{IJ} + A_{IJ}\,, \tag{13.65}$$

where S_{IJ} and A_{IJ} are the symmetric and antisymmetric parts of R_{IJ}, respectively. The symmetric part S_{IJ} can in fact be decomposed further:

$$S_{IJ} = \Big(S_{IJ} - \frac{1}{D-2}\delta_{IJ}S\Big) + \frac{1}{D-2}\,\delta_{IJ}\,S\,, \quad S \equiv S^{II} = \delta^{IJ}S_{IJ}\,. \tag{13.66}$$

The first term on the right-hand side is traceless:

$$\delta^{IJ}\Big(S_{IJ} - \frac{1}{D-2}\delta_{IJ}\,S\Big) = S - \frac{1}{D-2}\,\delta_{IJ}\,\delta^{IJ}S = 0\,, \tag{13.67}$$

since $\delta_{IJ}\,\delta^{IJ} = D-2$. Therefore (13.66) decomposes S_{IJ} into a traceless matrix plus a multiple of the unit matrix. Let $\hat{S}_{IJ}$ denote the traceless part of S_{IJ} and let $S' = S/(D-2)$. All in all, we have decomposed R_{IJ} as

$$R_{IJ} = \hat{S}_{IJ} + A_{IJ} + S'\delta_{IJ}\,. \tag{13.68}$$

This is the standard decomposition of a matrix into a symmetric-traceless part, an antisymmetric part, and a trace part. Each of the three pieces can be specified independently when writing a general matrix R_{IJ}. Therefore, the states in (13.64) can be split into three groups of linearly independent states:

$$\sum_{I,J} \hat{S}_{IJ}\,a_1^{I\dagger}\bar{a}_1^{J\dagger}|p^+, \vec{p}_T\rangle\,, \tag{13.69}$$

$$\sum_{I,J} A_{IJ}\,a_1^{I\dagger}\bar{a}_1^{J\dagger}|p^+, \vec{p}_T\rangle\,, \tag{13.70}$$

$$S'\,a_1^{I\dagger}\bar{a}_1^{I\dagger}\,|p^+, \vec{p}_T\rangle\,. \tag{13.71}$$

We now make a remarkable claim: the states (13.69) represent one-particle graviton states! We examined one-particle graviton states in Section 10.6. In the quantum theory of the free gravitational field these states are given by (10.110):

$$\sum_{I,J=2}^{D-1} \xi_{IJ}\,a^{IJ\dagger}_{p^+,\,p_T}|\Omega\rangle\,, \tag{13.72}$$

where ξ_{IJ} is an arbitrary symmetric traceless matrix. Since $\hat{S}_{IJ}$ is also a symmetric traceless matrix, the identification of states is possible if we identify the basis states:

$$a_1^{I\dagger}\bar{a}_1^{J\dagger}|p^+, \vec{p}_T\rangle \;\longleftrightarrow\; a^{IJ\dagger}_{p^+,\,p_T}|\Omega\rangle\,. \tag{13.73}$$

This identification is possible because the two sets of states have the same Lorentz labels, they carry the same momentum, and they have the same mass (both zero). This shows that the closed string has graviton states. Gravity has appeared in string theory! We never put in a dynamical metric and we never spoke about general covariance, yet somehow, the quantum states of the gravitational field have emerged!

The set of states in (13.70) corresponds to the one-particle states of the Kalb–Ramond field, an antisymmetric tensor field $B_{\mu\nu}$ with two indices. The light-cone analysis of this field was discussed in Problem 10.6 (see, in particular, parts (e) and (f)). The $B_{\mu\nu}$ field is

in many ways the tensor generalization of the Maxwell gauge field A_μ. The Kalb–Ramond field couples to strings in a way that is analogous to the way that the Maxwell field couples to particles. Thus, as we will see in Chapter 16, strings carry Kalb–Ramond charge.

There is one state that remains to be examined. The oscillator part of (13.71) has no free indices (I is summed over), so it represents one state. It corresponds to a one-particle state of a massless scalar field. This field is called the *dilaton*.

The above discussion of particle states is supplemented by an analysis of wavefunctions and field equations. Such analysis follows closely the treatment in Section 12.7. Wavefunctions $\psi_{IJ}(\tau, p^+, p_T)$ describe the general time-dependent states at the massless level of the closed string:

$$|\Psi, \tau\rangle = \int dp^+ d\vec{p}_T \;\psi_{IJ}(\tau, p^+, \vec{p}_T)\, a_1^{I\dagger} \tilde{a}_1^{J\dagger} |p^+, \vec{p}_T\rangle \,. \tag{13.74}$$

The Schrödinger equation satisfied by the states is $i\partial_\tau |\Psi, \tau\rangle = H|\Psi, \tau\rangle$. Using (13.50) and noting that for the states in question $N^\perp = \tilde{N}^\perp = 1$, we find

$$i\frac{\partial \psi_{IJ}}{\partial \tau} = \frac{\alpha'}{2}\, p^K p^K\, \psi_{IJ} \,. \tag{13.75}$$

The wavefunctions $\psi_{IJ}(\tau, p^+, p_T)$ become the fields of the classical field theories, with the Schrödinger equations interpreted as classical field equations. The symmetric-traceless part of ψ_{IJ} becomes the graviton field, the antisymmetric part becomes the Kalb–Ramond field, and the trace part becomes the dilaton field. The Schrödinger equations for the wavefunctions of the graviton states, the Kalb–Ramond states, and the dilaton state, are all included in (13.75), and can be separated by selecting the symmetric-traceless components, the antisymmetric components, and the trace component of ψ_{IJ}. On the other hand, in light-cone coordinates the massless scalar field equation $\partial^2\phi = 0$ takes the form (10.30):

$$\left(i\,\frac{\partial}{\partial x^+} - \frac{1}{2p^+}\, p^K p^K \right) \phi(x^+, p^+, \vec{p}_T) = 0 \,. \tag{13.76}$$

Setting $x^+ = \alpha' p^+ \tau$, we find

$$\left(i\,\frac{\partial}{\partial \tau} - \frac{\alpha'}{2}\, p^K p^K \right) \phi(\tau, p^+, \vec{p}_T) = 0 \,. \tag{13.77}$$

This equation takes the same form as (13.75). In fact, in the light-cone gauge, graviton fields, Kalb–Ramond fields, and the dilaton field, all satisfy the simple equation $\partial^2 \phi^{\cdots} = 0$, where the dots refer to the relevant indices. This is manifestly true for the massless dilaton, which is a scalar. For graviton fields, it was demonstrated in equation (10.107). For Kalb–Ramond fields, the equation $p^2 B^{\mu\nu} = 0$ was shown to hold in the light-cone gauge as part of the analysis in Problem 10.6.

In summary, at the massless level of the closed string we found gravity fields, Kalb–Ramond fields, and dilaton fields. Each of these fields deserves intense study. The gravity field is studied in general relativity. We will focus on the Kalb–Ramond field in Chapter 16.

The dilaton is a massless scalar field with surprising properties. A proper study of the dilaton belongs to an advanced course, but in the following section we discuss the role that it plays in string theory.

13.4 String coupling and the dilaton

The massless scalar field called the dilaton has a fascinating property: its expectation value controls the string coupling! This coupling is a dimensionless number that sets the strength of string interactions.

The classic example of a coupling constant is the electromagnetic fine-structure constant $\alpha \equiv e^2/(4\pi\hbar c) \simeq 1/137$. This dimensionless coupling constant controls the strength of electromagnetic interactions. In the hydrogen atom Hamiltonian, for example, α appears in the term which gives the electrostatic interaction energy between the proton and the electron. The mass-scale in the hydrogen atom is set by the mass m of the electron. Physical, dimensionful quantities, such as the binding energy E of the ground state, depend on the dimensionful parameter m of the theory, the fundamental constants $\hbar$ and c, *and* the dimensionless coupling α:

$$E = \frac{e^2}{4\pi}\frac{1}{2a_0} = \frac{1}{2}\Big(\frac{e^2}{4\pi\hbar c}\Big)^2 mc^2 = \frac{1}{2}\,\alpha^2\,(mc^2)\,, \tag{13.78}$$

where $a_0 = 4\pi\hbar^2/me^2$ is the Bohr radius. In the hypothetical limit $\alpha \to 0$, the binding energy vanishes, and the Bohr radius is infinite. This is what becomes of the hydrogen atom when we turn off electromagnetic interactions.

In string theory the story is not so different, at first. The dimensionful parameter can be taken to be α', which defines the string length $\ell_s = \sqrt{\alpha'}$ (working with $\hbar = c = 1$). Let g denote the dimensionless coupling for closed string interactions. If g was set to zero, then strings would not interact. Interactions in gravity are determined by the value of Newton's gravitational constant. If closed strings do not interact, gravitation would emerge without interactions, and the value of Newton's constant in string theory would be zero. Since it vanishes when $g \to 0$, Newton's constant is naturally expected to be proportional to some positive power of g; it turns out that it is proportional to g^2. Dimensional analysis fixes the α' dependence of Newton's constant. Equation (3.108) shows that the D-dimensional Newton constant $G^{(D)}$ has (natural) units of L^{D-2}. In fact, in natural units, $G^{(D)}$ is equal to the D-dimensional Planck length $\ell_{\rm P}^{(D)}$ to the power $(D-2)$. So $[G^{(26)}] = L^{24}$, and since $[\alpha'] = L^2$, we find

$$G^{(26)} \sim g^2\,(\alpha')^{12}\,. \tag{13.79}$$

Most phenomenological studies of string theory begin with ten-dimensional superstring theories. These theories contain both bosonic and fermionic excitations, so they include

the two types of particles that we observe in nature. The ten-dimensional Newton constant $G^{(10)}$ in superstring theory is given by

$$G^{(10)} = \left(\ell_{\rm P}^{(10)}\right)^8 \sim g^2\,(\alpha')^4\,. \tag{13.80}$$

It then follows that

$$\ell_{\rm P}^{(10)} \sim g^{1/4}\sqrt{\alpha'} = g^{1/4}\,\ell_s\,. \tag{13.81}$$

If the string coupling g is a small number, the string length is larger than the Planck length. If g is of order unity, the string length and the Planck length are comparable. To find four-dimensional implications, we use the relation (3.116) between higher- and lower-dimensional Newton constants in a compactification. Assume that six of the dimensions of the ten-dimensional world are curled up into a space with volume $V^{(6)}$. Then, the four-dimensional Newton constant G is related to the ten-dimensional one by

$$G = \frac{G^{(10)}}{V^{(6)}} \sim g^2\,\alpha'\,\frac{1}{V^{(6)}/(\alpha')^3}\,. \tag{13.82}$$

The ratio $V^{(6)}/(\alpha')^3$ is a dimensionless number that is typically assumed to be large. For a compactification of fixed volume, as measured in units of the string length, the four-dimensional Newton constant behaves like

$$G \sim g^2\,\alpha'\,. \tag{13.83}$$

In a theory with both open and closed strings, the open string coupling g_o is actually determined in terms of the closed string coupling g. One can prove that

$$g_o^2 \sim g\,. \tag{13.84}$$

This relation arises due to certain topological properties of two-dimensional world-sheets.

The coupling which controls the strength of an interaction may sometimes fail to be constant. Consider adding to a free Hamiltonian H_0 an interaction term $gH_{\rm int}$ that is proportional to a dimensionless coupling g. If g is declared to be a constant, you must specify its value by hand in order to define the complete Hamiltonian $H_0 + gH_{\rm int}$. But suppose that g is not a constant but rather a dynamical variable $g(t)$, and that the full Hamiltonian includes an additional term H_g that gives dynamics to g. In this case you cannot specify the coupling g arbitrarily by hand. The coupling would be determined, perhaps uniquely, or perhaps not, by the Hamiltonian equations derived from $H_0 + gH_{\rm int} + H_g$. If $g(t)$ is uniquely determined, no choice is needed. If $g(t)$ is not uniquely determined, some other criterion may be needed to select the physically realized solution.

The situation in string theory is similar; the string coupling can fail to be a constant. The closed string coupling g is *determined* by the value of the dilaton field $\phi(x)$:

$$g \sim e^{\phi}\,. \tag{13.85}$$

It follows that, in principle, the string coupling g is not an adjustable parameter of string theory. Rather, it is a *dynamical* parameter – a field, in fact. This is an ideal property in a

unified theory of all interactions, for it holds the promise that the string coupling may be calculable. On the other hand, it seems clear that the string coupling is not uniquely determined in string theory. Depending on the values taken by the other fields in the theory, the dilaton field may evolve in different ways. Under certain circumstances, the dilaton expectation value may even be an adjustable constant. One attractive possibility is that the other fields in the theory generate a potential for the dilaton. If this potential has a stable critical point the dilaton may be set equal to the critical value. Even more, the dilaton field would then acquire a mass. This is necessary for a realistic model of physics because there are no known massless scalars in nature.

If the string coupling is small, the quantum mechanical amplitudes for string interactions can be calculated accurately using known results about Riemann surfaces. The fascinating properties of Riemann surfaces allow one to understand why the infinities that plague quantum amplitudes in general relativity do not appear in string theory. This subject will be discussed in Chapter 26.

13.5 Closed strings on the $\mathbb{R}^1/\mathbb{Z}_2$ orbifold

We introduced orbifolds in §2.8 and stated that they are nontrivial spaces on which string propagation is tractable. In this and the following section we demonstrate this claim for the case of closed strings on the simplest orbifold, the half-line $\mathbb{R}^1/\mathbb{Z}_2$. Since we must deal with the full quantum string theory, the orbifold direction is just one direction in the 26-dimensional spacetime. This direction $x^{25} \equiv x$, is effectively restricted to $x \geq 0$ by the $\mathbb{Z}_2$ identification

$$x \sim -x\,. \tag{13.86}$$

No other coordinate is affected. The orbifold theory does not require boundary conditions at $x = 0$. The orbifold theory is defined by imposing a natural restriction on states of closed strings that live on the spacetime before orbifolding – a restriction on the spectrum of the original *parent* theory.

Writing $X^{25}(\tau,\sigma) \equiv X(\tau,\sigma)$, the collection of string coordinates is X^+, X^-, X^i, with $i = 2,\ldots,24$, and X. We introduce an operator U that implements on the string coordinates the $\mathbb{Z}_2$ transfomation that defines the orbifold. Thus, acting on X we must have

$$UX(\tau,\sigma)U^{-1} = -X(\tau,\sigma)\,. \tag{13.87}$$

Since U should not transform any other string coordinate we demand

$$UX^i(\tau,\sigma)U^{-1} = X^i(\tau,\sigma)\,. \tag{13.88}$$

Moreover, we demand the invariance of p^+ and x_0^-:

$$Up^+U^{-1} = p^+\,, \quad \text{and} \quad Ux_0^-U^{-1} = x_0^-\,, \tag{13.89}$$

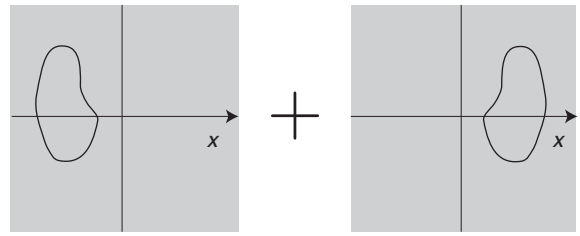

Fig. 13.1 **A closed string state invariant under $x \to -x$ is obtained by superposition of two states that go into each other as $x \to -x$.**

since these, together with (13.87) and (13.88), imply the invariance of X^+ and X^-:

$$U\, X^{\pm}(\tau,\sigma)\, U^{-1} = X^{\pm}(\tau,\sigma)\,. \tag{13.90}$$

Indeed, the invariance of $X^+ = \alpha' p^+ \tau$ follows directly from the invariance of p^+. The invariance of X^- holds because it is the sum of the invariant mode x_0^-, terms quadratic in the invariant coordinates, and terms quadratic in $\dot{X}$, X' (see (13.35)). It follows that p^- is invariant under the U action. Since p^+ is also invariant, the Hamiltonian $H = \alpha' p^+ p^-$ (see (13.49)) is invariant: $UHU^{-1} = H$. This means that U is a symmetry of the closed string theory.

The *orbifold* closed string theory keeps only U-invariant states of the parent theory. Since U is a symmetry, the truncation to U-invariant states is a consistent reduction: if the Hamiltonian were not U invariant, states that are U invariant at one time need not remain U invariant for all times. Intuitively, for U-invariant string states, the physics at $-x$ is determined by the physics at x, thus effectively making half of the space irrelevant. U-invariant states are naturally constructed by quantum superposition (Figure 13.1).

To implement the restriction to U-invariant states it is convenient to determine the action of U on the oscillators. The coordinate $X(\tau,\sigma)$ has the usual mode expansion (13.24) of a closed string coordinate

$$X(\tau,\sigma) = x_0 + \alpha' p\tau + i\sqrt{\frac{\alpha'}{2}} \sum_{n\neq 0} \frac{e^{-in\tau}}{n} \left(\alpha_n e^{in\sigma} + \bar{\alpha}_n e^{-in\sigma}\right)\,. \tag{13.91}$$

If we are to have equation (13.87) for all values of τ and σ we must have

$$\begin{aligned} U\, x_0\, U^{-1} &= -x_0\,, \quad U\, p\, U^{-1} = -p\,, \\ U\, \alpha_n\, U^{-1} &= -\alpha_n\,, \quad U\, \bar{\alpha}_n\, U^{-1} = -\bar{\alpha}_n\,. \end{aligned} \tag{13.92}$$

All operators in the expansion of X change sign under the action of U. All modes in the expansion of the coordinates X^i are left invariant by the action of U.

Quick calculation 13.3 Justify $UHU^{-1} = H$ directly from the oscillator expansion of $H = L_0^{\perp} + \bar{L}_0^{\perp} - 2$.

Let us now discuss the states of the theory, beginning with the ground states. We denote the ground states of the parent theory by $|p^+, \vec{p}, p\rangle$, where $\vec{p}$ is a vector with components

p^i and p denotes the momentum in the x direction. We also assume that states that have no momentum in the x direction are invariant under the action of U:

$$U|p^+, \vec{p}, 0\rangle = |p^+, \vec{p}, 0\rangle\,. \tag{13.93}$$

Being invariant under U, the state $|p^+, \vec{p}, 0\rangle$ is a ground state of the orbifold theory, but there are many more. To find these, we first derive the U action on $|p^+, \vec{p}, p\rangle$. We assume that for every pair of conjugate coordinate and momentum operators (q, p) momentum states are defined to satisfy $|p + \delta p\rangle = \exp(i\delta p\ q)|p\rangle$, with δp a constant. Applied to our pair (x_0, p), we find $|p^+, \vec{p}, p\rangle = e^{ix_0 p}|p^+, \vec{p}, 0\rangle$ and, therefore,

$$U|p^+, \vec{p}, p\rangle = Ue^{ix_0 p}U^{-1}U|p^+, \vec{p}, 0\rangle = e^{i(-x_0)p}|p^+, \vec{p}, 0\rangle = |p^+, \vec{p}, -p\rangle, \tag{13.94}$$

namely,

$$U|p^+, \vec{p}, p\rangle = |p^+, \vec{p}, -p\rangle\,. \tag{13.95}$$

It is now easy to form U-invariant states by linear combinations:

$$\text{Orbifold ground states:}\quad |p^+, \vec{p}, p\rangle + |p^+, \vec{p}, -p\rangle. \tag{13.96}$$

More precisely, orbifold ground states are general time-dependent superpositions of the above states:

$$\int \psi(\tau, p^+, \vec{p},\ p)\,|p^+, \vec{p}, p\rangle\, dp^+ d\vec{p}\ dp\,, \quad \psi(\tau, p^+, \vec{p}, -p) = \psi(\tau, p^+, \vec{p},\ p)\,. \tag{13.97}$$

The Fourier transformed ground state wavefunctions $\psi(\tau, p^+, \vec{p},\ x)$ are even functions of x. The massless states of the orbifold theory require $N^\perp = \bar{N}^\perp = 1$. Indeed, orbifolding does not change the mode expansion of X and the formula for M^2 is not changed. To build the states we need two oscillators, one barred one unbarred, acting on suitable ground states. Basis massless states are given by

$$\begin{aligned}
&\alpha^i_{-1}\bar{\alpha}^j_{-1}\left(|p^+, \vec{p}, p\rangle + |p^+, \vec{p}, -p\rangle\right),\\
&\alpha^i_{-1}\bar{\alpha}_{-1}\left(|p^+, \vec{p}, p\rangle - |p^+, \vec{p}, -p\rangle\right),\\
&\alpha_{-1}\bar{\alpha}^i_{-1}\left(|p^+, \vec{p}, p\rangle - |p^+, \vec{p}, -p\rangle\right),\\
&\alpha_{-1}\bar{\alpha}_{-1}\left(|p^+, \vec{p}, p\rangle + |p^+, \vec{p}, -p\rangle\right).
\end{aligned}$$

All these states are U-invariant. For those that have an odd number of X oscillators we use a combination of vacuum states with $U = -1$ so that the full state has $U = +1$. Massive states can be constructed similarly. It would seem that this is the end of the story, but there is a surprise. The orbifold theory includes more states than the ones discussed above. It contains a *twisted* sector with a new kind of closed strings. Let us examine these strings now.

13.6 The twisted sector of the orbifold

The additional closed strings that appear in the so-called twisted sector can be imagined as *open* strings in the parent theory, but with endpoints at locations identified by the orbifold condition (13.86). More precisely, we write

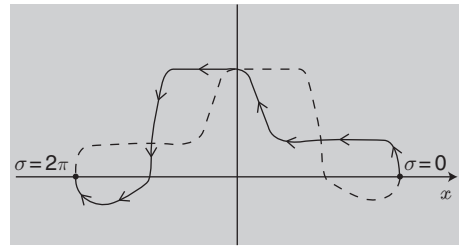

Fig. 13.2 **A "twisted" closed string satisfies $X(\tau, \sigma + 2\pi) = -X(\tau, \sigma)$. The string for $\sigma \in [0, 2\pi]$ is shown as a continuous line and the part $\sigma \in [2\pi, 4\pi]$ as a dashed line. The vertical coordinate represents any x^i.**

$$X(\tau, \sigma + 2\pi) = -X(\tau, \sigma)\,. \tag{13.98}$$

This equation tells us that $X(\tau, 2\pi) = -X(\tau, 0)$, which means that the orbifold identification makes them the same point and the string becomes effectively closed. For arbitrary σ the equation tells us that the string effectively repeats its position when σ is increased by 2π. In the parent space the string only closes after $\sigma \to \sigma + 4\pi$, as is clear from two applications of (13.98). A two-dimensional representation of a string obeying (13.98) is shown in Figure 13.2. We note that a twisted closed string must go through $x = 0$. This is clear because the continuous function $X(\tau, \sigma)$ in (13.98) takes both positive and negative values.

To develop the quantum theory of this sector we first find the appropriate oscillator expansion for the coordinate X that satisfies (13.98). As usual, to solve the wave equation, we write

$$X(\tau, \sigma) = X_L(u) + X_R(v)\,, \qquad u = \tau + \sigma\,,\ v = \tau - \sigma\,. \tag{13.99}$$

The constraint (13.98) implies that

$$X_L(u + 2\pi) + X_R(v - 2\pi) = -X_L(u) - X_R(v), \tag{13.100}$$

or, equivalently,

$$X_L(u + 2\pi) + X_L(u) = -\big(X_R(v) + X_R(v - 2\pi)\big). \tag{13.101}$$

Taking derivatives with respect to u and v we find

$$X_L'(u + 2\pi) = -X_L'(u)\,, \qquad X_R'(v + 2\pi) = -X_R'(v). \tag{13.102}$$

For an ordinary string coordinate $X_L^{i\,\prime}$ and $X_R^{i\,\prime}$ are periodic functions with period 2π. For X the derivatives reverse sign when the argument changes by 2π. To write a mode expansion that is conveniently normalized we mimic (13.14). In order to get a sign change when $u \to u + 2\pi$ we need exponentials of the form $\exp(iku)$ with k half-integer. We therefore write

$$X'_L(u) = \sqrt{\frac{\alpha'}{2}} \sum_{n\in\mathbb{Z}_{\text{odd}}} \bar{\alpha}_{\frac{n}{2}}\, e^{-i\frac{n}{2}u}\,,$$
$$X'_R(v) = \sqrt{\frac{\alpha'}{2}} \sum_{n\in\mathbb{Z}_{\text{odd}}} \alpha_{\frac{n}{2}}\, e^{-i\frac{n}{2}v}\,. \tag{13.103}$$

The labels on the oscillators, now half-integers, match the factors of u or v in the exponentials. Integrating (13.103) we get

$$X_L(u) = x_L + i\sqrt{\frac{\alpha'}{2}} \sum_{n\in\mathbb{Z}_{\text{odd}}} \frac{2}{n}\,\bar{\alpha}_{\frac{n}{2}}\, e^{-i\frac{n}{2}u}$$
$$X_R(v) = x_R + i\sqrt{\frac{\alpha'}{2}} \sum_{n\in\mathbb{Z}_{\text{odd}}} \frac{2}{n}\,\alpha_{\frac{n}{2}}\, e^{-i\frac{n}{2}v}\,. \tag{13.104}$$

Equation (13.101) now gives $x_L = -x_R$, or $x_L + x_R = 0$. Putting the left- and right-moving parts together we arrive at the mode expansion for the twisted sector closed strings:

$$X(\tau,\sigma) = i\sqrt{\frac{\alpha'}{2}} \sum_{n\in\mathbb{Z}_{\text{odd}}} \frac{2}{n}\, e^{-i\frac{n}{2}\tau} \left(\bar{\alpha}_{\frac{n}{2}} e^{-i\frac{n}{2}\sigma} + \alpha_{\frac{n}{2}} e^{i\frac{n}{2}\sigma}\right). \tag{13.105}$$

The expansion contains neither coordinate nor momentum zero modes. It costs energy to get away from $x = 0$ since at least one point on the string must remain at $x = 0$. As we will see later, the fields associated with twisted states are localized at the fixed point $x = 0$ and depend only on x^+, x^-, and x^i.

The commutation relations for the modes can be derived in complete analogy to the usual case. The coordinate obeys $[X(\tau,\sigma)\,, \mathcal{P}^\tau(\tau,\sigma')] = i\delta(\sigma - \sigma')$. Moreover, given our mode expansions the coordinate X satisfies (13.26) with $n \to n/2$ and sums over odd numbers:

$$\dot{X} + X' = 2X'_L(\tau+\sigma) = \sqrt{2\alpha'} \sum_{n\in\mathbb{Z}_{\text{odd}}} \bar{\alpha}_{\frac{n}{2}}\, e^{-i\frac{n}{2}(\tau+\sigma)}\,,$$
$$\dot{X} - X' = 2X'_R(\tau-\sigma) = \sqrt{2\alpha'} \sum_{n\in\mathbb{Z}_{\text{odd}}} \alpha_{\frac{n}{2}} e^{-i\frac{n}{2}(\tau-\sigma)}\,. \tag{13.106}$$

Applying the top-sign version of (13.28) to X, we get

$$\sum_{m',n'\in\mathbb{Z}_{\text{odd}}} e^{-i\frac{m'}{2}(\tau+\sigma)} e^{-i\frac{n'}{2}(\tau+\sigma')} \left[\bar{\alpha}_{\frac{m'}{2}}\,, \bar{\alpha}_{\frac{n'}{2}}\right] = 2\pi\, i\, \frac{d}{d\sigma}\delta(\sigma - \sigma')\,. \tag{13.107}$$

To extract the commutators we apply the following integral operators to the left and right sides of equation (13.107):

$$\frac{1}{2\pi}\int_0^{2\pi} d\sigma\, e^{i\frac{m}{2}\sigma} \;\cdot\; \frac{1}{2\pi}\int_0^{2\pi} d\sigma'\, e^{i\frac{n}{2}\sigma'}\,, \quad m, n \in \mathbb{Z}_{\text{odd}}\,. \tag{13.108}$$

Two functions $e^{i\frac{k}{2}\sigma}$ and $e^{i\frac{k'}{2}\sigma}$ with $k, k' \in \mathbb{Z}_{\text{odd}}$ and $k + k' \neq 0$ are orthogonal over the interval $[0, 2\pi]$ because $k + k'$ is an even integer. The integrations in (13.108) therefore select a single commutator on the left-hand side, and one can show that

$$\left[\bar{\alpha}_{\frac{m}{2}}\,, \bar{\alpha}_{\frac{n}{2}}\right] = \tfrac{m}{2}\,\delta_{m+n,0}\,. \tag{13.109}$$

This is the expected form of the commutation relations.

Quick calculation 13.4 Prove (13.109).

Similar commutation relations hold for the right-moving oscillators and, as usual, left-moving and right-moving oscillators commute:

$$\left[\alpha_{\frac{m}{2}}, \alpha_{\frac{n}{2}}\right] = \tfrac{m}{2}\delta_{m+n,0}\,, \quad \left[\alpha_{\frac{m}{2}}, \bar{\alpha}_{\frac{n}{2}}\right] = 0\,. \tag{13.110}$$

We must still implement the orbifold identification, so equation (13.87) must hold for the twisted X in (13.105). We can readily read off the action of U on the new oscillators:

$$U\alpha_{\frac{n}{2}}U^{-1} = -\alpha_{\frac{n}{2}}\,, \qquad U\bar{\alpha}_{\frac{n}{2}}U^{-1} = -\bar{\alpha}_{\frac{n}{2}}\,. \tag{13.111}$$

The absence of a momentum zero mode in the expansion of X means that twisted states do not have a conserved momentum along the X direction. U-invariant ground states are labeled by p^+ and the transverse momentum $\vec{p}$ along the X^i:

$$\text{twisted sector ground states:} \quad |p^+, \vec{p}\,\rangle\,. \tag{13.112}$$

These are not to be confused with the ground states (13.96) in the "untwisted" sector that have zero momentum along X.

To discuss excited states in the twisted sector we need the relevant formula for $\alpha' M^2$. This time we can expect changes because the half-integer moding of the X oscillators can alter the ordering constant. To find out, we recall that in open string theory we anticipated the value of the ordering constant by performing the naive ordering of $L_0^\perp$ and using the ζ-function motivated rule $1 + 2 + 3 + \cdots \to -\frac{1}{12}$. Let us do the same now for the closed string operator $\bar{L}_0^\perp$. We see from the top equation in (13.36) that $\bar{L}_0^\perp$ is determined by contributions from $(\dot{X}^I + X^{I\prime})^2$, that is, $(\dot{X}^i + X^{i\prime})^2 + (\dot{X} + X')^2$. Since the mode expansions of $\dot{X}^i + X^{i\prime}$ and those of $\dot{X} + X'$ are completely analogous, the formula for $\bar{L}_0^\perp$ in (13.37) will be modified to read

$$\bar{L}_0^\perp = \frac{1}{2}\sum_{p\in\mathbb{Z}} \bar{\alpha}_p^i \bar{\alpha}_{-p}^i + \frac{1}{2}\sum_{k\in\mathbb{Z}_{\text{odd}}} \bar{\alpha}_{\frac{k}{2}}\bar{\alpha}_{-\frac{k}{2}}\,. \tag{13.113}$$

Each of the twenty-three directions i in the first sum contributes an ordering constant of $\frac{1}{2}\cdot(-\frac{1}{12}) = -\frac{1}{24}$. The new contribution arises from ordering the second summand:

$$\begin{aligned}\frac{1}{2}\sum_{k\in\mathbb{Z}_{\text{odd}}}\bar{\alpha}_{\frac{k}{2}}\bar{\alpha}_{-\frac{k}{2}} &= \frac{1}{2}\sum_{k\in\mathbb{Z}^+_{\text{odd}}}\bar{\alpha}_{\frac{k}{2}}\bar{\alpha}_{-\frac{k}{2}} + \frac{1}{2}\sum_{k\in\mathbb{Z}^+_{\text{odd}}}\bar{\alpha}_{-\frac{k}{2}}\bar{\alpha}_{\frac{k}{2}} \\ &= \sum_{k\in\mathbb{Z}^+_{\text{odd}}}\bar{\alpha}_{-\frac{k}{2}}\bar{\alpha}_{\frac{k}{2}} + \frac{1}{2}\sum_{k\in\mathbb{Z}^+_{\text{odd}}}\left[\bar{\alpha}_{\frac{k}{2}}\,,\bar{\alpha}_{-\frac{k}{2}}\right] \\ &= \sum_{k\in\mathbb{Z}^+_{\text{odd}}}\bar{\alpha}_{-\frac{k}{2}}\bar{\alpha}_{\frac{k}{2}} + \frac{1}{4}\sum_{k\in\mathbb{Z}^+_{\text{odd}}}k\ . \end{aligned} \tag{13.114}$$

To evaluate this we need to calculate the "sum" of all the odd positive integers. This is done as follows:

$$\sum_{k=1}^{\infty}k = \sum_{k\in\mathbb{Z}^+_{\text{odd}}}k + \sum_{k\in\mathbb{Z}^+_{\text{even}}}k = \sum_{k\in\mathbb{Z}^+_{\text{odd}}}k + 2\sum_{k=1}^{\infty}k\,. \tag{13.115}$$

It then follows that

$$\sum_{k\in\mathbb{Z}^+_{\text{odd}}}k = -\sum_{k=1}^{\infty}k = \frac{1}{12}\,. \tag{13.116}$$

We can now write the precise quantum form of (13.113):

$$\bar{L}_0^{\perp} = \frac{1}{4}\alpha'\,p^i\,p^i + \sum_{p=1}^{\infty}\bar{\alpha}^i_{-p}\bar{\alpha}^i_p - 23\cdot\frac{1}{24} + \sum_{k\in\mathbb{Z}^+_{\text{odd}}}\bar{\alpha}_{-\frac{k}{2}}\bar{\alpha}_{\frac{k}{2}} + \frac{1}{4}\cdot\frac{1}{12}\,, \tag{13.117}$$

which we write as

$$\bar{L}_0^{\perp} = \frac{1}{4}\alpha'\,p^i\,p^i + \bar{N}^{\perp} - \frac{15}{16}\,,\quad \bar{N}^{\perp} = \sum_{p=1}^{\infty}\bar{\alpha}^i_{-p}\bar{\alpha}^i_p + \sum_{k\in\mathbb{Z}^+_{\text{odd}}}\bar{\alpha}_{-\frac{k}{2}}\bar{\alpha}_{\frac{k}{2}}\,. \tag{13.118}$$

A completely analogous formula holds for $L_0^{\perp}$.

Quick calculation 13.5 Show that $[\bar{N}^{\perp},\bar{\alpha}_{-\frac{q}{2}}] = \frac{q}{2}\bar{\alpha}_{-\frac{q}{2}}$ and explain why $\bar{N}^{\perp}$ is properly called a number operator.

To write the mass-squared formula recall that $\frac{1}{2}\alpha' M^2$ is equal to the sum of $N^{\perp}$ plus $\bar{N}^{\perp}$ plus the ordering constants, as in (13.48). We thus have

$$\frac{1}{2}\alpha' M^2 = N^{\perp} + \bar{N}^{\perp} - \frac{15}{8}\,. \tag{13.119}$$

The twisted sector ground states $|p^+,\vec{p}\,\rangle$ have $N^{\perp} = \bar{N}^{\perp} = 0$. They are tachyonic states with $\frac{1}{2}\alpha' M^2 = -\frac{15}{8}$. The first excited states are built using the lowest moded twisted oscillators:

$$\alpha_{-\frac{1}{2}}\bar{\alpha}_{-\frac{1}{2}}|p^+,\vec{p}\rangle\,,\qquad \frac{1}{2}\alpha' M^2 = \frac{1}{2}+\frac{1}{2}-\frac{15}{8} = -\frac{7}{8}. \tag{13.120}$$

The momentum labels on twisted states indicate that the wavefunctions are of the form $\psi(\tau, p^+, \vec{p})$, or $\psi(\tau, x^-, \vec{x})$ in coordinate space. Given the correspondence between wavefunctions and fields, we conclude that the fields associated with twisted states have no x argument. These fields must live at some specific value of x. Given that energetics forces twisted states to localize around $x = 0$, we reasonably conclude that the fields live on the reduced 25-dimensional spacetime defined by $x = 0$. The above ground states and first excited states are associated with scalar fields because the states have no indices along directions in the reduced spacetime.

Orbifolds always have a twisted sector thus orbifolding has a double effect – the so-called *double strike*. One effect is that states of the parent theory that are not invariant under the orbifold action are projected out. The other effect is that we gain a whole new sector of states that satisfy "twisted" boundary conditions.

Problems

Problem 13.1 Commutation relations for oscillators.

(a) Use the lower-sign version of equation (13.28) and the appropriate mode expansion to verify explicitly that the unbarred commutation relations of (13.29) emerge.

(b) The set of functions $e^{in\sigma}$, with $n \in \mathbb{Z}$, is complete on the interval $\sigma \in [0, 2\pi]$. Use this fact to prove that

$$\delta(\sigma - \sigma') = \frac{1}{2\pi} \sum_{n \in \mathbb{Z}} e^{in(\sigma - \sigma')} . \tag{1}$$

(c) Compute explicitly the commutator $[X^I(\tau, \sigma), \mathcal{P}^{\tau J}(\tau, \sigma')]$ using the mode expansions of X and $\mathcal{P}$ and the commutation relations (13.29), (13.30), and (13.33). Use equation (1) to confirm that the expected answer (13.27) emerges.

(d) Prove the zero mode commutation relations (13.33), starting with a derivation of

$$\left[x_0^I + \sqrt{2\alpha'}\alpha_0^I \tau \, , \; \dot{X}^J(\tau, \sigma') \right] = i\alpha' \eta^{IJ} ,$$

which is the closed string analog of equation (12.47).

Problem 13.2 A projector into physical states.

Consider the vector space $\mathcal{H}$ spanned by the set of states $|\lambda, \bar{\lambda}\rangle$ in equation (13.60). Explain why, for any state $|\lambda, \bar{\lambda}\rangle \in \mathcal{H}$, the eigenvalue of $P = L_0^\perp - \bar{L}_0^\perp$ is an integer. Show that

$$\mathcal{P}_0 = \int_0^{2\pi} \frac{d\theta}{2\pi} \, e^{-i(L_0^\perp - \bar{L}_0^\perp)\theta} ,$$

is a projector from $\mathcal{H}$ into the vector subspace where $P = 0$. Thus $\mathcal{P}_0$ projects into the state space of closed strings.

Problem 13.3 Action of $L_0^\perp - \bar{L}_0^\perp$.

(a) Prove that equation (13.59) holds for finite σ_0. You may find it useful to define $f(\sigma_0) = e^{-iP\sigma_0}\, X^I(\tau,\sigma)\, e^{iP\sigma_0}$ and to calculate multiple derivatives of f, evaluated at $\sigma_0 = 0$.

(b) Explain why

$$e^{-iP\sigma_0}\,(\dot{X}^I \pm X^{I\prime})(\tau,\sigma)\,e^{iP\sigma_0} = (\dot{X}^I \pm X^{I\prime})(\tau,\sigma+\sigma_0)\,. \tag{1}$$

(c) Use equation (1) to calculate $e^{-iP\sigma_0}\,\alpha_n^I\,e^{iP\sigma_0}$ and $e^{-iP\sigma_0}\,\bar{\alpha}_n^I\,e^{iP\sigma_0}$. In doing so, you are finding the action of a σ translation on the oscillators.

(d) Consider the state

$$|U\rangle = \alpha_{-m}^I \bar{\alpha}_{-n}^J |p^+, \vec{p}_T\rangle\,, \quad m,n>0\,.$$

Use the results of (c) to calculate $e^{-iP\sigma_0}|U\rangle$. What is the condition that makes the state $|U\rangle$ invariant under σ translations?

Problem 13.4 $L_0^\perp - \bar{L}_0^\perp$ as world-sheet momentum.

(a) Use equations (13.36) to show that

$$L_0^\perp - \bar{L}_0^\perp = -\frac{1}{2\pi\alpha'}\int_0^{2\pi} d\sigma\, \dot{X}^I X^{I\prime}\,. \tag{1}$$

Also prove that

$$L_0^\perp - \bar{L}_0^\perp = -\frac{p^+}{2\pi}\int_0^{2\pi} d\sigma\, \frac{\partial X^-}{\partial\sigma}\,,$$

which explains that $L_0^\perp - \bar{L}_0^\perp$ vanishes classically because X^-, just like any other string coordinate, must satisfy the closed string periodicity condition.

(b) The dynamics of the transverse light-cone string coordinates is governed by the Lagrangian density (12.81):

$$\mathcal{L} = \frac{1}{4\pi\alpha'}\left(\dot{X}^I\,\dot{X}^I - X^{I\prime}\,X^{I\prime}\right).$$

Show that the infinitesimal constant σ translation $\delta X^I = \epsilon\,\partial_\sigma X^I$ is a symmetry of $\mathcal{L}$ in the sense of Problem 8.10. Calculate the associated charge and show that it is proportional to $L_0^\perp - \bar{L}_0^\perp$, as given in (1).

Problem 13.5 Unoriented closed strings.

This problem is the closed string version of Problem 12.12. The closed string $X^\mu(\tau,\sigma)$ with $\sigma \in [0, 2\pi]$ and fixed τ is a parameterized closed curve in spacetime. The orientation of a string is the direction of increasing σ on this curve.

(a) Consider now the closed string $X^\mu(\tau, 2\pi - \sigma)$ with the same τ as above. How is this second string related to the first string above? How are their orientations related? Make a rough sketch, showing the original string as a continuous line and the second string as a dashed line.

Introduce an orientation reversing *twist* operator Ω such that

$$\Omega\, X^I(\tau,\sigma)\,\Omega^{-1} = X^I(\tau, 2\pi - \sigma)\,. \tag{1}$$

Moreover, declare that

$$\Omega\, x_0^-\,\Omega^{-1} = x_0^-\,, \quad \Omega\, p^+\,\Omega^{-1} = p^+\,. \tag{2}$$

(b) Use the closed string oscillator expansion (13.24) to calculate

$$\Omega\, x_0^I\,\Omega^{-1}\,, \quad \Omega\,\alpha_0^I\,\Omega^{-1}\,, \quad \Omega\,\alpha_n^I\,\Omega^{-1}\,, \quad \text{and} \quad \Omega\,\tilde{\alpha}_n^I\,\Omega^{-1}\,.$$

(c) Show that $\Omega\, X^-(\tau,\sigma)\,\Omega^{-1} = X^-(\tau, 2\pi - \sigma)$. Since $\Omega\, X^+(\tau,\sigma)\,\Omega^{-1} = X^+(\tau, 2\pi - \sigma)$, equation (1) actually holds for all string coordinates. We say that orientation reversal is a symmetry of closed string theory because it leaves the closed string Hamiltonian H invariant: $\Omega H \Omega^{-1} = H$. Explain why this is true.

(d) Assume that the ground states are twist invariant. List the closed string states for $N^\perp \leq 2$, and give their twist eigenvalues. If you were commissioned to build a theory of *unoriented* closed strings, which of the states would you have to discard? What are the massless fields of unoriented closed string theory?

Problem 13.6 Orientifold Op-planes.

An orientifold Op-plane is a a hyperplane with p spatial dimensions, just as a Dp-brane has p spatial dimensions. The Op-plane arises when we perform a truncation which keeps only the closed string states that are invariant under the symmetry transformation which simultaneously reverses the string orientation and reflects the coordinates normal to the Op-plane.

For an Op-plane, let $x^1, \ldots, x^p$ be the directions along the Op-plane, and let $x^{p+1}, \ldots, x^d$ with $d = 25$ be the directions orthogonal to the Op-plane. The Op-plane position is defined by $x^a = 0$ for $a = p+1, \ldots, d$. We will organize the string coordinates as $X^+, X^-, \{X^i\}, \{X^a\}$ with $i = 2, \ldots, p,$ and $a = p+1, \ldots, d$. Let Ω_p denote the operator generating the transformation

$$\Omega_p\, X^a(\tau,\sigma)\,\Omega_p^{-1} = -X^a(\tau, 2\pi - \sigma)\,, \tag{1}$$

$$\Omega_p\, X^i(\tau,\sigma)\,\Omega_p^{-1} = X^i(\tau, 2\pi - \sigma)\,. \tag{2}$$

Moreover, assume that

$$\Omega_p\, x_0^-\,\Omega_p^{-1} = x_0^-\,, \quad \Omega_p\, p^+\,\Omega_p^{-1} = p^+\,. \tag{3}$$

(a) For an O23-plane the two normal directions x^{24}, x^{25} can be represented by a plane. A closed string at a fixed τ appears as a parameterized closed curve $X^a(\tau,\sigma)$ in this plane. Draw such an oriented closed string that lies fully in the first quadrant of the (x^{24}, x^{25}) plane. Draw also the string $\tilde{X}^a(\tau,\sigma) = -X^a(\tau, 2\pi - \sigma)$.

(b) For any operator $\mathcal{O}$ the action of Ω_p is defined by $\mathcal{O} \to \Omega_p \mathcal{O} \Omega_p^{-1}$. Use the expansion (13.24) to calculate the action of Ω_p on the operators

$$x_0^a,\ p^a,\ \alpha_n^a,\ \bar{\alpha}_n^a, \quad x_0^i,\ p^i,\ \alpha_n^i, \quad \text{and} \quad \bar{\alpha}_n^i\,.$$

Show that $\Omega_p\, X^{\pm}(\tau,\sigma)\, \Omega_p^{-1} = X^{\pm}(\tau, 2\pi - \sigma)$. Explain why the orientifold transformations are symmetries of closed string theory.

(c) Denote the ground states by $|p^+, p^i, p^a\rangle$. Assume that the states $|p^+, p^i, \vec{0}\,\rangle$ are invariant under Ω_p: $\Omega_p|p^+, p^i, \vec{0}\,\rangle = |p^+, p^i, \vec{0}\,\rangle$. Prove that

$$\Omega_p|p^+, p^i, p^a\,\rangle = |p^+, p^i, -p^a\,\rangle\,.$$

Assume that, for every pair of conjugate coordinate and momentum operators (q, p), the momentum states are defined to satisfy $|p + \delta p\rangle = \exp(i\delta p\ q)|p\rangle$, with δp a constant.

The massless states of the oriented closed string theory are given by

$$|\Phi\rangle = \int dp^+ d\vec{p}^{\,i}\, d\vec{p}^{\,a}\, \Phi^{\pm}_{IJ}(\tau, p^+, p^i, p^a) \left(\alpha^I_{-1}\bar{\alpha}^J_{-1} \pm \alpha^J_{-1}\bar{\alpha}^I_{-1}\right) |p^+, p^i, p^a\rangle,$$

where $\Phi^{\pm}_{IJ}$ are wavefunctions, and the transverse light-cone indices I, J run over all the values taken by the indices i and a.

Now consider the truncation to Ω_p invariant states: $\Omega_p|\Phi\rangle = |\Phi\rangle$. The resulting string theory is the theory in the presence of the orientifold Op-plane. Intuitively, for an invariant state the amplitudes for the string to lie along each of the two related curves of part (a) are equal.

(d) Find the conditions that must be satisfied by $\Phi^{\pm}_{ab}$, $\Phi^{\pm}_{ia}$, $\Phi^{\pm}_{ij}$ to guarantee Ω_p invariance. All of the conditions are of the form

$$\Phi^{\pm}_{IJ}(\tau, p^+, p^i, p^a) = \cdots\, \Phi^{\pm}_{IJ}(\tau, p^+, p^i, -p^a)\,,$$

where the dots represent a sign factor that you must determine for each case.

Remarks: in coordinate space, letting $x^m = \{x^0, \ldots, x^p\}$, the above invariance conditions require that

$$\Phi^{\pm}_{IJ}(x^m, x^a) = \cdots\, \Phi^{\pm}_{IJ}(x^m, -x^a)\,.$$

In the presence of an orientifold plane, the values of the fields at (x^m, x^a) determine the values of the fields at the reflected point $(x^m, -x^a)$. The fields are either even or odd under $x^a \to -x^a$. The orientifold plane is some kind of mirror that relates the physics at reflected points, effectively cutting the space in half. In one half of the space, and away from the orientifold, there are no constraints, so one has the full set of fields of oriented closed strings. An O25-plane is space filling. Since it has no normal directions, the orientifold symmetry includes only string orientation reversal. This case was studied in Problem 13.5.

14 A look at relativistic superstrings

Realistic string theories must contain fermionic states, like the states of electrons or quarks. Superstrings include anticommuting dynamical variables in addition to the commuting coordinates X^μ that describe the position of strings. For open superstrings, quantization gives a state space with a Neveu–Schwarz (NS) sector that contains bosonic states and a Ramond (R) sector that contains fermionic states. The theory has supersymmetry, a symmetry that ensures that the number of bosonic and fermionic degrees of freedom are the same at any mass level. We examine type II closed string theories, which arise by tensoring the state spaces of open superstrings.

14.1 Introduction

We have so far studied bosonic string theories, both open and closed. These string theories live in 26-dimensional spacetime, and all of their quantum states represent bosonic particle states. Among them we found important bosonic particles, such as the photon and the graviton. Non-Abelian gauge bosons, needed to transmit the strong and weak forces, also arise in bosonic string theory, as we will see in Chapter 15.

Realistic string theories, however, must also contain the states of fermionic particles. You may recall that a quantum state of identical bosonic particles is symmetric under the exchange of any two of the particles. A quantum state of identical fermionic particles, on the other hand, is antisymmetric under the exchange of any two of the particles. Quarks and leptons are fermionic particles. To obtain them we need *superstring* theories. We will not study superstrings in detail in this book. A proper explanation would require a detailed discussion of spinors and the Dirac equation in various numbers of dimensions, as well as other technicalities. This would take us too long. Here we give you a brief discussion of the basics of superstrings, enough to appreciate that they are natural generalizations of the bosonic string with some interesting new ingredients. Certain applications that are discussed in this book involve superstrings; this chapter provides the required background material.

The superstring spectrum contains no tachyon. As a result, the theory of open superstrings can describe D-branes that are stable. Since bosonic string theory D-branes are always unstable, this affords new and interesting possibilities. In addition, superstrings have supersymmetry. Supersymmetry is a symmetry that relates the bosonic and fermionic quantum states of the theory: in a superstring theory we find an equal number of fermionic

and bosonic states at every mass level. If supersymmetry exists in nature it must be spontaneously broken: we do not observe degeneracies between fermions and bosons. Many physicists believe that supersymmetry is an attractive candidate for the physics that lies beyond the presently established Standard Model of particle physics.

14.2 Anticommuting variables and operators

We describe the position of classical bosonic strings using the string coordinates $X^\mu(\tau, \sigma)$. The $X^\mu(\tau, \sigma)$ are classical *commuting* variables: products of them are independent of the order of the factors. In the quantum theory the X^μs become operators that do not generally commute. The failure of two operators A and B to commute is encoded in the commutator $[A, B] \equiv AB - BA$.

To get fermions in string theory we introduce new dynamical world-sheet variables $\psi_1^\mu(\tau, \sigma)$ and $\psi_2^\mu(\tau, \sigma)$. The classical variables ψ_α^μ ($\alpha = 1, 2$) are not ordinary commuting variables, but rather *anticommuting* variables. Since this is an important concept let us digress.

Two variables b_1 and b_2 are said to anticommute with one another if

$$b_1 b_2 = -b_2 b_1. \tag{14.1}$$

If b_1 and b_2 are anticommuting variables, more is true. The variables must anticommute with themselves. Thus $b_1 b_1 = -b_1 b_1$, which means that

$$b_1 b_1 = 0, \qquad b_2 b_2 = 0. \tag{14.2}$$

For classical anticommuting variables the order of factors is important. If we have a set of anticommuting variables b_i indexed by i, then we have

$$b_i b_j = -b_j b_i, \quad \forall i, j. \tag{14.3}$$

Quick calculation 14.1 Verify that the matrices γ^1 and γ^2 defined by

$$\gamma^1 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \qquad \gamma^2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{14.4}$$

anticommute, but are not anticommuting variables.

In a quantum theory classical anticommuting variables become quantum operators that can sometimes fail to anticommute. Given two such quantum operators f_1 and f_2, the failure to anticommute is measured by the *anticommutator* $\{f_1, f_2\}$, defined by

$$\{f_1, f_2\} \equiv f_1 f_2 + f_2 f_1. \tag{14.5}$$

If two quantum operators anticommute, their anticommutator is zero.

We can explain what anticommuting operators have to do with fermions. Recall that the quantization of a scalar field, for example, gave us creation and annihilation operators

that create particles. The creation operators $a_p^\dagger$ all commute among themselves and help us construct the multiparticle states

$$(a_{p_1}^\dagger)^{n_1}\,(a_{p_2}^\dagger)^{n_2}\ldots(a_{p_k}^\dagger)^{n_k}|\Omega\rangle, \tag{14.6}$$

that contain n_1 particles with momentum $\vec{p}_1$, n_2 particles with momentum $\vec{p}_2$, and so on. The n_is are arbitrary positive integers.

To describe the relativistic electron and its antiparticle, the positron, one uses the classical Dirac field, a classical anticommuting field variable. The quantization of the Dirac field gives rise to creation and annihilation operators for electrons and different creation and annihilation operators for positrons. For convenience, let us focus on electrons. The creation operators $f_{p,s}^\dagger$ are labeled by momentum $\vec{p}$ and spin s. Electrons are spin one-half particles, so the spin label can only take two values. All the electron creation operators are anticommuting variables and they all anticommute. In particular, this means that $f_{p,s}^\dagger f_{p,s}^\dagger = 0$, for any $\vec{p}$ and s. A multiparticle state of electrons takes the form

$$f_{p_1,s_1}^\dagger f_{p_2,s_2}^\dagger\ldots f_{p_k,s_k}^\dagger|\Omega\rangle \tag{14.7}$$

and describes a state with an electron with momentum $\vec{p}_1$ and spin s_1, an electron with momentum $\vec{p}_2$ and spin s_2, and so on. Note that we *cannot* get a state with two electrons that have the same momentum and the same spin because $f_{p,s}^\dagger f_{p,s}^\dagger|\Omega\rangle = 0$. The anticommuting creation operators automatically implement Fermi's exclusion principle. In fact, if we use a wavefunction to create a superposition of states

$$\sum_{s_1,s_2}\int d\vec{p}\;\psi\big(p_1,s_1;\,p_2,s_2\big)\,f_{p_1,s_1}^\dagger f_{p_2,s_2}^\dagger|\Omega\rangle, \tag{14.8}$$

we can quickly conclude that only the part of the wavefunction ψ that is antisymmetric under the simultaneous exchange $p_1 \leftrightarrow p_2$ and $s_1 \leftrightarrow s_2$ contributes to the above state.

Quick calculation 14.2 Prove the above claim.

14.3 World-sheet fermions

As we mentioned before, classical superstrings require anticommuting dynamical variables $\psi_\alpha^\mu(\tau,\sigma)$, with $\alpha = 1, 2$. Recall that for each value of μ the dynamical variable $X^\mu(\tau,\sigma)$ is a world-sheet boson. As it turns out, for each μ, the two components ψ_1^μ and ψ_2^μ comprise the variables needed to describe a fermion on the (τ,σ) world, that is, a *world-sheet* fermion. Remarkably, the quantization of such objects results in particle states that behave as *spacetime* fermions, which is what we need.

In light-cone quantization X^+ was set proportional to τ and X^- was solved for in terms of other quantities. With superstrings this remains true but, in addition, the light-cone gauge condition sets $\psi_\alpha^+ = 0$ and allows one to solve for ψ_α^-. Both X^- and ψ_α^- receive contributions from the transverse X^I and ψ_α^I. Since both X^μ and ψ_α^μ are spacetime Lorentz

vectors, both enter into the definition of the light-cone Lorentz generator M^{-I}. It follows that the commutator $[M^{-I}, M^{-J}]$ receives contributions from both X^μ and ψ^μ_α. The requirement that this commutator vanishes gives constraints different from those obtained for bosonic strings. The number of spacetime dimensions is no longer twenty-six, but rather ten. The downward shift of the mass-squared equals minus one-half, rather than minus one.

In the light-cone gauge we need only concern ourselves with the transverse fields $\psi^I_\alpha(\tau,\sigma)$. To study their quantization it is convenient to have an action, denoted as S_ψ, that describes their dynamics. This action is the fermion analog of the action (12.81) that describes the dynamics of the transverse coordinates X^I. So, all in all, the action S that describes the full set of degrees of freedom takes the form

$$S = \frac{1}{4\pi\alpha'}\int d\tau \int_0^\pi d\sigma \Big(\dot{X}^I\,\dot{X}^I - X^{I\prime}\,X^{I\prime}\Big) + S_\psi, \tag{14.9}$$

with

$$S_\psi = \frac{1}{2\pi}\int d\tau \int_0^\pi d\sigma \Big[\psi^I_1(\partial_\tau+\partial_\sigma)\psi^I_1 + \psi^I_2(\partial_\tau-\partial_\sigma)\psi^I_2\Big]. \tag{14.10}$$

The action S_ψ is the Dirac action for a fermion that lives in the two-dimensional (τ,σ) world. Note that each term in S_ψ contains just one derivative. For bosons, like the X^I, terms in the action contain two derivatives. A term in the Lagrangian that couples a field to itself and has just one derivative risks being an irrelevant total derivative. Indeed, consider an arbitrary field h and a coupling $h(\partial_\tau h)$. We have

$$h(\partial_\tau h) = \partial_\tau(hh) - (\partial_\tau h)h, \tag{14.11}$$

or, equivalently,

$$h(\partial_\tau h) + (\partial_\tau h)h = \partial_\tau(hh). \tag{14.12}$$

If the field h is commuting, the two terms on the left-hand side are identical and, indeed, $h(\partial_\tau h)$ is a total derivative. If the field h is anticommuting, $(\partial_\tau h)h = -h(\partial_\tau h)$ and both the left-hand side and the right-hand side of the equation vanish. We learn that $h(\partial_\tau h)$, for h anticommuting, is not a total derivative. This means that the action S_ψ is nontrivial only because the ψ^I_α fields are anticommuting.

Let us vary the fields ψ^I_α in S_ψ in order to find the equations of motion and the boundary conditions. We have

$$\begin{aligned}\delta S_\psi = \frac{1}{2\pi}\int d\tau \int_0^\pi d\sigma \Big[&\delta\psi^I_1(\partial_\tau+\partial_\sigma)\psi^I_1 + \psi^I_1(\partial_\tau+\partial_\sigma)\delta\psi^I_1 \\ &+ \delta\psi^I_2(\partial_\tau-\partial_\sigma)\psi^I_2 + \psi^I_2(\partial_\tau-\partial_\sigma)\delta\psi^I_2\Big].\end{aligned} \tag{14.13}$$

Consider the second term on the first line of the above right-hand side:

$$\begin{aligned}\psi^I_1(\partial_\tau+\partial_\sigma)\delta\psi^I_1 &= \partial_\tau(\psi^I_1\,\delta\psi^I_1) + \partial_\sigma(\psi^I_1\,\delta\psi^I_1) - [(\partial_\tau+\partial_\sigma)\psi^I_1]\delta\psi^I_1 \\ &= \partial_\tau(\psi^I_1\,\delta\psi^I_1) + \partial_\sigma(\psi^I_1\,\delta\psi^I_1) + \delta\psi^I_1(\partial_\tau+\partial_\sigma)\psi^I_1.\end{aligned} \tag{14.14}$$

A similar calculation can be done for the second term on the second line of (14.13). Back into this equation and, as usual, ignoring the total derivatives along time,

$$\delta S_\psi = \frac{1}{\pi}\int d\tau \int_0^\pi d\sigma \Big[\, \delta\psi_1^I(\partial_\tau + \partial_\sigma)\psi_1^I + \delta\psi_2^I(\partial_\tau - \partial_\sigma)\psi_2^I \Big] + \frac{1}{2\pi}\int d\tau \Big[\psi_1^I \delta\psi_1^I - \psi_2^I \delta\psi_2^I\Big]_{\sigma=0}^{\sigma=\pi} . \tag{14.15}$$

We can now read immediately the equations of motion

$$(\partial_\tau + \partial_\sigma)\psi_1^I = 0, \qquad (\partial_\tau - \partial_\sigma)\psi_2^I = 0, \tag{14.16}$$

as well as the boundary conditions

$$\psi_1^I(\tau,\sigma_*)\,\delta\psi_1^I(\tau,\sigma_*) - \psi_2^I(\tau,\sigma_*)\,\delta\psi_2^I(\tau,\sigma_*) = 0, \tag{14.17}$$

that must hold both at $\sigma_* = 0$ and at $\sigma_* = \pi$ for all times.

We will use the above equations of motion and boundary conditions to discuss the quantization and the state space of the theory. Our analysis will not be complete. In particular, we will not discuss the anticommutation relations satisfied by the fields ψ_α^I and, as a result, we will not justify the anticommutation relations of the corresponding oscillators. Nevertheless, the discussion will illuminate the main features of the quantization, and all results should seem plausible.

We begin by noting that the equations of motion (14.16) imply that ψ_1^I is right-moving and ψ_2^I is left-moving:

$$\psi_1^I(\tau,\sigma) = \Psi_1^I(\tau-\sigma),$$
$$\psi_2^I(\tau,\sigma) = \Psi_2^I(\tau+\sigma). \tag{14.18}$$

Let us now consider the boundary conditions (14.17). What can we do to satisfy these conditions? A little thought shows that attempts to make each term in (14.17) vanish do not work. Try, for example, setting $\psi_1^I(\tau,0) = 0$. Our solution in (14.18) then implies $\Psi_1^I(\tau) = 0$, for all τ, which results in $\psi_1^I(\tau,\sigma) \equiv 0$.

The situation improves if we impose boundary conditions that relate ψ_1^I and ψ_2^I at the endpoints. For each σ_* we take

$$\psi_1^I(\tau,\sigma_*) = \pm\psi_2^I(\tau,\sigma_*), \tag{14.19}$$

where the choice of sign is still to be determined. If the fields are so constrained at an endpoint, their variations must also respect this condition:

$$\delta\psi_1^I(\tau,\sigma_*) = \pm\delta\psi_2^I(\tau,\sigma_*). \tag{14.20}$$

It then follows by multiplication of the last two equations that (14.17) holds for either choice of sign.

Let us now discuss the signs. Since both ψ_1^I and ψ_2^I appear quadratically in the action, their signs can be changed without physical consequence. This arbitrariness is used conventionally to demand that at the endpoint $\sigma_* = 0$,

$$\psi_1^I(\tau,0) = \psi_2^I(\tau,0). \tag{14.21}$$

Since we cannot change the sign of ψ_1^I or ψ_2^I without changing this condition, the choice of sign at the other endpoint is physically relevant:

$$\psi_1^I(\tau,\pi) = \pm\psi_2^I(\tau,\pi). \tag{14.22}$$

The full superstring theory state space breaks into two subspaces or, as they are typically called, two *sectors*: a Ramond (R) sector which contains the states that arise using the top choice of sign, and a Neveu–Schwarz (NS) sector which contains the states that arise using the lower choice of sign. The boundary conditions are better understood by assembling a fermion field Ψ^I defined over the full interval $\sigma \in [-\pi,\pi]$:

$$\Psi^I(\tau,\sigma) \equiv \begin{cases} \psi_1^I(\tau,\,\sigma) & \sigma \in [\,0,\,\pi\,] \\ \psi_2^I(\tau,-\sigma) & \sigma \in [-\pi,0]. \end{cases} \tag{14.23}$$

This construction is reminiscent of (12.36), which defined a field over $\sigma \in [-\pi,\pi]$ for bosonic open strings. The boundary condition (14.21) guarantees that Ψ^I is continuous at $\sigma = 0$. Moreover, on account of (14.18), $\Psi^I(\tau,\sigma)$ is a function of $\tau - \sigma$:

$$\Psi^I(\tau,\sigma) = \chi^I(\tau-\sigma). \tag{14.24}$$

Finally, the boundary condition (14.22) gives

$$\Psi^I(\tau,\pi) = \psi_1^I(\tau,\pi) = \pm\psi_2^I(\tau,\pi) = \pm\Psi^I(\tau,-\pi). \tag{14.25}$$

We thus learn that a periodic fermion Ψ^I corresponds to Ramond boundary conditions and an antiperiodic fermion Ψ^I corresponds to Neveu–Schwarz boundary conditions:

$$\begin{aligned} &\Psi^I(\tau,\pi) = +\Psi^I(\tau,-\pi) \quad \text{Ramond boundary condition,} \\ &\Psi^I(\tau,\pi) = -\Psi^I(\tau,-\pi) \quad \text{Neveu–Schwarz boundary condition.} \end{aligned} \tag{14.26}$$

Let us consider both cases in detail now.

14.4 Neveu–Schwarz sector

Since the Neveu–Schwarz fermion Ψ^I is a function of $\tau - \sigma$ and changes sign when $\sigma \to \sigma + 2\pi$, it must be expanded with fractionally moded exponentials:

$$\Psi^I(\tau,\sigma) \sim \sum_{r\in\mathbb{Z}+1/2} b_r^I\, e^{-ir(\tau-\sigma)}, \tag{14.27}$$

up to a normalization factor that will not enter our discussion. Indeed, for any $r = n + \frac{1}{2}$, with n integer, we have

$$e^{ir(\sigma+2\pi)} = e^{ir\sigma}e^{i\left(n+\frac{1}{2}\right)2\pi} = e^{ir\sigma}e^{i\pi} = -e^{ir\sigma}, \tag{14.28}$$

which guarantees that Ψ^I is antiperiodic. Since Ψ^I is anticommuting, the expansion coefficients b_r^I are anticommuting operators. Following our usual notation, the negatively moded coefficients $b_{-1/2}^I$, $b_{-3/2}^I$, $b_{-5/2}^I$, …, are creation operators, while the positively

moded ones $b^I_{1/2}, b^I_{3/2}, b^I_{5/2}, \ldots$ are annihilation operators. These operators act on a vacuum, henceforth called the Neveu–Schwarz vacuum $|\text{NS}\rangle$. These operators satisfy the anticommutation relation

$$\{b^I_r, b^J_s\} = \delta_{r+s,0}\delta^{IJ}. \tag{14.29}$$

Given that all creation operators anticommute with one another and square to zero, each b^I_{-r} can appear at most once in any state. Since the $X^I(\tau,\sigma)$ are quantized as usual, we still have the α^I_{-n} creation operators. Thus, the states in the NS sector are of the form:

$$\text{NS sector: } |\lambda\rangle = \prod_{I=2}^{9}\prod_{n=1}^{\infty}(\alpha^I_{-n})^{\lambda_{n,I}}\prod_{J=2}^{9}\prod_{r=\frac{1}{2},\frac{3}{2},\ldots}(b^J_{-r})^{\rho_{r,J}}|\text{NS}\rangle\otimes|p^+,\vec{p}_T\rangle. \tag{14.30}$$

Here the $\rho_{r,J}$ are either zero or one. We have written the full ground state as a "product" $\otimes$ of the ground state $|\text{NS}\rangle$ for the b^J_{-r} operators and the ground state $|p^+,\vec{p}_T\rangle$ for the α^I_{-n} operators. The order in which the b operators appear in the state does not matter when we consider a single state. Since all the bs anticommute, different orderings can only differ by an overall sign, and no new states are obtained.

The mass-squared operator in the NS sector, before normal ordering, is given by

$$M^2 = \frac{1}{\alpha'}\Big(\frac{1}{2}\sum_{p\neq 0}\alpha^I_{-p}\alpha^I_p + \frac{1}{2}\sum_{r\in\mathbb{Z}+\frac{1}{2}} r b^I_{-r}\, b^I_r\Big). \tag{14.31}$$

We can use the heuristic method based on ζ-functions to find the ordering constant "a" in M^2 – the constant that must be added inside the above parenthesis when the sums are replaced by their normal-ordered versions. For the bosonic α^I oscillators we know that each coordinate contributes $-\frac{1}{24}$ to a. We record this piece of information as

$$a_B = -\frac{1}{24}. \tag{14.32}$$

For the NS fermions the terms in M^2 that require reordering are

$$\begin{aligned}\frac{1}{2}\sum_{r=-\frac{1}{2},-\frac{3}{2},\ldots} r b^I_{-r}\, b^I_r &= \frac{1}{2}\sum_{r=\frac{1}{2},\frac{3}{2},\ldots}(-r) b^I_r\, b^I_{-r}\\ &= \frac{1}{2}\sum_{r=\frac{1}{2},\frac{3}{2},\ldots} r b^I_{-r}\, b^I_r - \frac{1}{2}(D-2)\Big(\frac{1}{2}+\frac{3}{2}+\frac{5}{2}+\cdots\Big).\end{aligned} \tag{14.33}$$

In the first step we let $r\to -r$ and in the second step we used the anticommutator (14.29). The sum over odd positive integers was evaluated in (13.116) and gives $+\frac{1}{12}$. It then follows that

$$\frac{1}{2}\sum_{r=-\frac{1}{2},-\frac{3}{2},\ldots} r b^I_{-r}\, b^I_r = \frac{1}{2}\sum_{r=\frac{1}{2},\frac{3}{2},\ldots} r b^I_{-r}\, b^I_r - \frac{1}{48}(D-2). \tag{14.34}$$

We have thus learned that each NS fermion (antiperiodic fermion) contributes to a the constant a_{NS} given by

$$a_{\text{NS}} = -\frac{1}{48}. \tag{14.35}$$

The full ordering constant for M^2 is therefore

$$a = (D-2)(a_B + a_{\rm NS}) = (D-2)\Big(-\frac{1}{24} - \frac{1}{48}\Big) = -(D-2)\frac{1}{16}. \tag{14.36}$$

For $D = 10$ we find $a = -\frac{1}{2}$ and therefore (14.31) gives

$$M^2 = \frac{1}{\alpha'}\big(-\tfrac{1}{2} + N^\perp\big), \quad \text{with} \quad N^\perp = \sum_{p=1}^{\infty} \alpha^I_{-p}\alpha^I_p + \sum_{r=\frac{1}{2},\frac{3}{2}\ldots} r\, b^I_{-r}\, b^I_r. \tag{14.37}$$

$N^\perp$ is a number operator that counts the contributions from the α^Is and the b^Is.

Quick calculation 14.3 To test that $N^\perp$ includes the fermionic contribution to the number, show that the eigenvalue of $N^\perp$ on $b^I_{-r_1} b^J_{-r_2}|{\rm NS}\rangle$, with $r_1, r_2 > 0$, is $r_1 + r_2$.

We can now list the first few levels of states in the NS sector. Organized by the number eigenvalue or, equivalently, mass-squared, we have

$$\begin{aligned}
&\alpha' M^2 = -\tfrac{1}{2},\ N^\perp = 0 : |{\rm NS}\rangle \otimes |p^+, \vec{p}_T\rangle,\\
&\alpha' M^2 = \ \ 0,\ N^\perp = \tfrac{1}{2} : b^I_{-1/2}|{\rm NS}\rangle \otimes |p^+, \vec{p}_T\rangle,\\
&\alpha' M^2 = \ \ \tfrac{1}{2},\ N^\perp = 1 : \big\{\alpha^I_{-1},\ b^I_{-1/2} b^J_{-1/2}\big\}|{\rm NS}\rangle \otimes |p^+, \vec{p}_T\rangle,\\
&\alpha' M^2 = 1,\ N^\perp = \tfrac{3}{2} : \big\{\alpha^I_{-1} b^J_{-1/2},\ b^I_{-3/2},\ b^I_{-1/2} b^J_{-1/2} b^K_{-1/2}\big\}|{\rm NS}\rangle \otimes |p^+, \vec{p}_T\rangle.
\end{aligned} \tag{14.38}$$

The states with $N^\perp = 0$ have $\alpha' M^2 = -\frac{1}{2}$. The states with $N^\perp = \frac{1}{2}$ are massless. There are eight of them, labeled by the vector index I.

Quick calculation 14.4 How many states are there at $N^\perp = \frac{3}{2}$?

It is useful to have an operator whose value on states is $+1$ if the state is bosonic and -1 if the state is fermionic. This operator is usually called $(-1)^F$, where F stands for fermion number. This is reasonable, states with even fermion number are bosonic and states with odd fermion number are fermionic. To calculate $(-1)^F$ on any state we must first give the eigenvalue of $(-1)^F$ on the Neveu–Schwarz ground states $|{\rm NS}\rangle \otimes |p^+, \vec{p}_T\rangle$. Let us declare that number to be minus one, thus making the ground states fermionic:

$$(-1)^F\, |{\rm NS}\rangle \otimes |p^+, \vec{p}_T\rangle = -\,|{\rm NS}\rangle \otimes |p^+, \vec{p}_T\rangle. \tag{14.39}$$

The eigenvalue of $(-1)^F$ on a state is equal to minus one times a sequence of factors of minus one, one for each fermionic oscillator that appears in the state. Thus, acting on the generic state (14.30) we get

$$(-1)^F\, |\lambda\rangle = -(-1)^{\sum_{r,J} \rho_{r,J}}\, |\lambda\rangle. \tag{14.40}$$

Operationally, this result follows if we take $(-1)^F$ to anticommute with all of the fermionic operators

$$\big\{(-1)^F, b^I_r\big\} = 0. \tag{14.41}$$

Consider again our list of states (14.38). Since fermionic oscillators contribute half-integers to $N^\perp$, states with integer $N^\perp$ must have an even number of fermionic oscillators. So all states with integer $N^\perp$ have $(-1)^F = -1$; they are fermionic states. All states with half-integer $N^\perp$ have an odd number of fermionic oscillators and therefore $(-1)^F = +1$. These are bosonic states, and they include the eight massless states on the second line of the table.

The fermion or boson character of the states we have discussed so far is restricted to the (τ, σ) world-sheet, where the Ψ fields are fermions. We will later discuss if the above states are fermions or bosons in spacetime.

14.5 Ramond sector

With Ramond boundary conditions (14.26) the field Ψ^I is periodic and can be expanded using integrally moded oscillators:

$$\Psi^I(\tau, \sigma) \sim \sum_{n\in\mathbb{Z}} d_n^I \, e^{-in(\tau-\sigma)}. \tag{14.42}$$

Since Ψ^I is anticommuting, the oscillators d_n^I are anticommuting operators. Again, the negatively moded oscillators $d_{-1}^I, d_{-2}^I, d_{-3}^I, \ldots$, are creation operators, while the positively moded ones $d_1^I, d_2^I, d_3^I, \ldots$ are annihilation operators. The Ramond oscillators satisfy the anticommutation relation

$$\{d_m^I, d_n^J\} = \delta_{m+n,0}\delta^{IJ}. \tag{14.43}$$

As in the NS sector, the Ramond creation operators are all anticommuting and, as a result, can appear at most only once on any given state.

Ramond fermions are more complicated than NS fermions because the eight fermionic zero modes d_0^I must be treated with care. It turns out that these eight operators can be organized by simple linear combinations into four creation operators and four annihilation operators. Let us call the four creation operators

$$\xi_1, \ \xi_2, \ \xi_3, \ \xi_4. \tag{14.44}$$

Being zero modes, these creation operators do not contribute to the mass-squared of the states. Postulating a unique vacuum $|0\rangle$, the creation operators allow us to construct $16 = 2^4$ degenerate Ramond ground states. In fact, eight of these states have an even number of ξs acting on $|0\rangle$ and the other eight have an odd number of ξs acting on $|0\rangle$. Explicitly, the eight states $|R_a\rangle$, $a = 1, 2, \ldots, 8$, with an even number of creation operators are

$$\text{states } |R_a\rangle : \quad \begin{cases} |0\rangle, \\ \xi_1\xi_2|0\rangle, \ \xi_1\xi_3|0\rangle, \ \xi_1\xi_4|0\rangle, \ \xi_2\xi_3|0\rangle, \ \xi_2\xi_4|0\rangle, \ \xi_3\xi_4|0\rangle, \\ \xi_1\xi_2\xi_3\xi_4|0\rangle. \end{cases} \tag{14.45}$$

The eight states $|R_{\bar{a}}\rangle$, $\bar{a} = \bar{1}, \bar{2}, \ldots, \bar{8}$, with an odd number of creation operators are

$$\text{states } |R_{\bar{a}}\rangle : \begin{cases} \xi_1|0\rangle,\ \xi_2|0\rangle,\ \xi_3|0\rangle,\ \xi_4|0\rangle, \\ \xi_1\xi_2\xi_3|0\rangle,\ \xi_1\xi_2\xi_4|0\rangle,\ \xi_1\xi_3\xi_4|0\rangle,\ \xi_2\xi_3\xi_4|0\rangle. \end{cases} \tag{14.46}$$

The states $|R_a\rangle$ and $|R_{\bar{a}}\rangle$ comprise the full set of degenerate Ramond ground states, denoted as $|\mathrm{R}_A\rangle$ with $A = 1, \ldots, 16$. The Ramond sector of the state space contains the states

$$\text{R sector:} \quad |\lambda\rangle = \prod_{I=2}^{9}\prod_{n=1}^{\infty}(\alpha_{-n}^{I})^{\lambda_{n,I}}\prod_{J=2}^{9}\prod_{m=1}^{\infty}(d_{-m}^{J})^{\rho_{m,J}}|\mathrm{R}_A\rangle \otimes |p^+, \vec{p}_T\rangle. \tag{14.47}$$

Here the $\rho_{m,J}$ are either zero or one.

Just like in the NS sector, the Ramond sector has an $(-1)^F$ operator. This operator anticommutes with all the fermion oscillators, including the zero modes:

$$\left\{(-1)^F,\ d_n^I\right\} = 0, \tag{14.48}$$

and, additionally, we conventionally declare $|0\rangle$ to be fermionic

$$(-1)^F|0\rangle = -|0\rangle. \tag{14.49}$$

It thus follows that all eight $|R_a\rangle$ states are fermionic and all $|R_{\bar{a}}\rangle$ states are bosonic.

The mass-squared operator in the R sector, before normal ordering, is given by

$$M^2 = \frac{1}{\alpha'}\Big(\frac{1}{2}\sum_{p\neq 0}\alpha_{-p}^I\alpha_p^I + \frac{1}{2}\sum_{n\in\mathbb{Z}} n d_{-n}^I\, d_n^I\Big). \tag{14.50}$$

For the R fermions the terms that require ordering are

$$\begin{aligned} \tfrac{1}{2}\sum_{n=-1,-2,\ldots} n d_{-n}^I\, d_n^I &= -\tfrac{1}{2}\sum_{n=1,2,\ldots} n d_n^I\, d_{-n}^I \\ &= \tfrac{1}{2}\sum_{n=1,2,\ldots} n d_{-n}^I\, d_n^I - \frac{1}{2}(D-2)(1+2+3+\cdots) \\ &= \tfrac{1}{2}\sum_{n=1,2,\ldots} n d_{-n}^I\, d_n^I + \frac{1}{24}(D-2). \end{aligned} \tag{14.51}$$

We thus learn that the ordering contribution from a (periodic) Ramond fermion is

$$a_{\mathrm{R}} = \frac{1}{24}. \tag{14.52}$$

This number is precisely the opposite of a_B. Since there are equal numbers of bosonic coordinates X^I and Ramond fermions, their respective normal-ordering contributions to the mass-squared cancel out and (14.50) gives:

$$M^2 = \frac{1}{\alpha'}\sum_{n\geq 1}\Big(\alpha_{-n}^I\alpha_n^I + n d_{-n}^I\, d_n^I\Big). \tag{14.53}$$

This formula implies that all the Ramond ground states are massless.

Let us now list the states at various mass levels

$$\begin{aligned}
\alpha' M^2 = 0: &\qquad |R_a\rangle \;\Big\|\; |R_{\tilde{a}}\rangle \\
\alpha' M^2 = 1: &\qquad \alpha^I_{-1}|R_a\rangle,\ d^I_{-1}|R_{\tilde{a}}\rangle \;\Big\|\; \alpha^I_{-1}|R_{\tilde{a}}\rangle,\ d^I_{-1}|R_a\rangle, \\
\alpha' M^2 = 2: &\quad \{\alpha^I_{-2},\ \alpha^I_{-1}\alpha^J_{-1}, d^I_{-1}d^J_{-1}\}|R_a\rangle, \Big\| \{\alpha^I_{-2},\ \alpha^I_{-1}\alpha^J_{-1}, d^I_{-1}d^J_{-1}\}|R_{\tilde{a}}\rangle, \\
&\qquad \{\alpha^I_{-1}d^J_{-1},\ d^I_{-2}\}|R_{\tilde{a}}\rangle \;\Big\|\; \{\alpha^I_{-1}d^J_{-1},\ d^I_{-2}\}|R_a\rangle. \qquad (14.54)
\end{aligned}$$

We have separated the states in two groups with identical number of states: to the left of the bars we find the states with $(-1)^F = -1$ (fermionic states) and to the right of the bars the states with $(-1)^F = +1$ (bosonic states). Note that for each state to the left there is a corresponding state to the right built on a ground state of opposite fermion number. The appearance of an equal number of bosonic and fermionic states at every mass level is a signal of supersymmetry. This is, however, supersymmetry on the world-sheet. As we will soon see, spacetime supersymmetry arises only after we combine states from the Ramond and Neveu–Schwarz sectors.

14.6 Counting states

Before assembling a supersymmetric theory we digress to learn how to count the number of states that a string theory has at any given mass level. Our goal is to obtain generating functions that encode these numbers for the NS and R sectors.

The generating functions contain the information about numbers of states in their power series expansions. Typically we want a function $f(x)$ such that

$$f(x) = \sum_{n=0}^{\infty} a(n)\, x^n, \qquad (14.55)$$

where $a(n)$ is the number of states with, say, $N^\perp = n$. Suppose we have just one oscillator $a_1^\dagger$. There is then just one state $|0\rangle$ with $N^\perp = 0$, one state $a_1^\dagger|0\rangle$ with $N^\perp = 1$ and, in fact, one state $(a_1^\dagger)^k|0\rangle$ with $N^\perp = k$. It follows that the function $f_1(x)$ corrresponding to this system is

$$f_1(x) = 1 + x + x^2 + x^3 + \cdots = \frac{1}{1-x}. \qquad (14.56)$$

Suppose we had just one oscillator $a_2^\dagger$ with mode number two. Then we get one vacuum and one state for each even value of $N^\perp$, resulting in a function $f_2(x)$ that takes the form:

$$f_2(x) = 1 + x^2 + x^4 + x^6 + \cdots = \frac{1}{1-x^2}. \qquad (14.57)$$

We now ask: what is the function $f_{12}(x)$ corresponding to the states built using both $a_1^\dagger$ and $a_2^\dagger$? To obtain these states we form all the products where the first factor is a state built with $a_1^\dagger$s and the second factor is a state built with $a_2^\dagger$s. In forming these products one removes

the ground states, multiplies the oscillators, and restores the ground state. It follows that f_{12} is given by the product of f_1 and f_2:

$$(1 + x + x^2 + x^3 + \cdots)(1 + x^2 + x^4 + x^6 + \cdots). \tag{14.58}$$

To aid the imagination one can place the oscillator content of the states on the generating functions:

$$(1 + a_1^\dagger x + (a_1^\dagger)^2 x^2 + (a_1^\dagger)^3 x^3 + \cdots)(1 + a_2^\dagger x^2 + (a_2^\dagger)^2 x^4 + (a_2^\dagger)^3 x^6 + \cdots).$$

It is clear that the product forms all possible states, and states with $N^\perp = k$ appear together with x^k. This makes it clear that (14.58) is the correct answer:

$$f_{12}(x) = f_1(x) f_2(x) = \frac{1}{1-x} \frac{1}{1-x^2}. \tag{14.59}$$

If we use oscillators $a_1^\dagger, a_2^\dagger, a_3^\dagger, \ldots$ of all mode numbers, the generating function is

$$f(x) = \prod_{n=1}^{\infty} \frac{1}{1-x^n} = \frac{1}{1-x} \frac{1}{1-x^2} \frac{1}{1-x^3} \cdots. \tag{14.60}$$

The above use of multiplication is quite general. Consider a state space with oscillators of type A and generating function $f_A(x)$ as well as a state space with oscillators of type B and generating function $f_B(x)$. Then, the generating function for the state space built with oscillators of types A and B is $f_{AB}(x) = f_A(x) f_B(x)$.

As an application we compute the generating function for bosonic open string theory. In this theory we have oscillators of all mode numbers, but they come in 24 species, one for each transverse light-cone direction. Since each species gives a generating function (14.60), the full generating function for bosonic open string theory is obtained by raising the right-hand side of (14.60) to the 24th power:

$$\prod_{n=1}^{\infty} \frac{1}{(1-x^n)^{24}}. \tag{14.61}$$

Since mass-squared is more physical that $N^\perp$ it is useful to use $\alpha' M^2$ instead of $N^\perp$ to define the generating function. In a generating function based on $\alpha' M^2$ the coefficient of x^k counts the number of states with $\alpha' M^2 = k$. For open bosonic strings $\alpha' M^2 = N^\perp - 1$, so the $\alpha' M^2$ generating function is obtained by dividing the $N^\perp$ generating function by one power of x. As a result, the generating function $f_{os}(x)$ for bosonic open string theory is

$$f_{os}(x) = \frac{1}{x} \prod_{n=1}^{\infty} \frac{1}{(1-x^n)^{24}}. \tag{14.62}$$

With the help of a symbolic manipulator we readily find that

$$f_{os}(x) = \frac{1}{x} + 24 + 324\,x + 3200\,x^2 + 25\,650\,x^3 + 176\,256\,x^4 + \cdots. \tag{14.63}$$

This equation reminds us that we have a tachyon with $\alpha' M^2 = -1$, 24 massless states of a Maxwell field, and 324 states with $\alpha' M^2 = +1$.

Quick calculation 14.5 Use the binomial formula to calculate, by hand, the terms in $f_{os}(x)$ up to and including $\mathcal{O}(x^2)$.

Quick calculation 14.6 Construct explicitly all the states with $\alpha' M^2 = 2$ and count them, verifying that there are indeed a total of 3200 states. You may find the counting formula in Problem 12.11 useful.

To obtain the generating functions for the NS and R sectors we must learn how to count states built with fermionic oscillators. Happily, this is easy to do. Again, we begin by using $N^\perp$ and assuming that we have a single fermionic creation operator f_{-r} that contributes r to $N^\perp$. For this oscillator we can only build two states $|0\rangle$ and $f_{-r}|0\rangle$. The generating function $f_r(x)$ is therefore

$$f_r(x) = 1 + x^r. \tag{14.64}$$

Since the NS sector contains oscillators $b^I_{-1/2}, b^I_{-3/2}, \ldots$ coming in eight species, the generating function associated with these is

$$\left[(1+x^{1/2})(1+x^{3/2})(1+x^{5/2})\cdots\right]^8 = \prod_{n=1}^{\infty}(1+x^{n-\frac{1}{2}})^8. \tag{14.65}$$

Recalling that $\alpha' M^2 = N^\perp - \frac{1}{2}$ and that we have eight bosonic coordinates as well, the $\alpha' M^2$ based generating function $f_{\rm NS}(x)$ for the NS sector is

$$f_{\rm NS}(x) = \frac{1}{\sqrt{x}}\prod_{n=1}^{\infty}\Big(\frac{1+x^{n-\frac{1}{2}}}{1-x^n}\Big)^8. \tag{14.66}$$

Expanding for the first few orders we get

$$f_{\rm NS}(x) = \frac{1}{\sqrt{x}} + 8 + 36\sqrt{x} + 128\,x + 402\,x\sqrt{x} + 1152\,x^2 + \cdots. \tag{14.67}$$

This expansion shows the tachyon at $\alpha' M^2 = -1/2$, the eight massless states, and the 36 states at $\alpha' M^2 = 1/2$. The corresponding states were listed in (14.38).

For the Ramond sector we have $\alpha' M^2 = N^\perp$, with no offset. Since the fermionic oscillators $d^I_{-1}, d^I_{-2}, \ldots$ are integrally moded we get

$$f_{\rm R}(x) = 16\prod_{n=1}^{\infty}\Big(\frac{1+x^n}{1-x^n}\Big)^8. \tag{14.68}$$

The overall multiplicative factor appears because each combination of oscillators gives rise to 16 states by acting on each of the available ground states. The power series expansion of (14.68) gives

$$f_{\rm R}(x) = 16 + 256\,x + 2304\,x^2 + 15360\,x^3 + \cdots. \tag{14.69}$$

The NS generating function contains both integer and half-integer powers of x, while the R generating function only has integer powers of x. We note that the R coefficients are

actually double the corresponding NS coefficients. This is not a coincidence, as we will see in the following section.

14.7 Open superstrings

We have seen that the Ramond sector has world-sheet supersymmetry: there are equal numbers of fermionic and bosonic states at each mass level. Consider, for example, the ground states. There are sixteen of them, obtained by acting on $|0\rangle$ with four linear combinations built from the zero modes d_0^I. The states break into two groups $|R_a\rangle$ and $|R_{\bar{a}}\rangle$ of eight states each, with opposite values of $(-1)^F$.

Since the d_0^I carry a Lorentz index they form a Lorentz vector and transform as such under Lorentz transformations. The ground states, however, are built in an intricate way using the zero modes (see (14.45) and (14.47)), so they do *not* transform as vectors. In fact, under Lorentz transformations the $|R_a\rangle$ transform among themselves and so do the $|R_{\bar{a}}\rangle$. Both transform as *spinors*, the kind of transformation that is appropriate for states that are *spacetime* fermions. Both the index a and the index $\bar{a}$ are spinor indices, but they label somewhat different spinors. This reflects the fact that there are two different kinds of fermions in a ten-dimensional spacetime. Curiously, if you have just one fermion there is no way of telling which kind it is, but once you have two of them, you can tell if they are of the same type or of different type.

So, do we get two spacetime fermions from the R sector ground states? There are two reasons to believe that the answer should be no. First, there is something strange about getting spacetime fermions from *both* $|R_a\rangle$ and $|R_{\bar{a}}\rangle$, given that these states have opposite values of $(-1)^F$ and thus rather different commuting character. Second, with two spacetime fermions we would not get spacetime supersymmetry. Identifying $|R_a\rangle$ as spacetime fermions and $|R_{\bar{a}}\rangle$ as spacetime bosons is not an alternative either, since spacetime bosons cannot carry a spinor index.

A strategy then emerges. Since all states in the R sector have a spinor index, we will only attempt to get spacetime fermions from this sector. We also recognize that all fermions must arise from states with the same value of $(-1)^F$. Following Gliozzi, Scherk, and Olive (GSO) we proceed to truncate the Ramond sector down to the set of states with $(-1)^F = -1$. These are the states to the left of the bars in (14.54). With our conventions, these are world-sheet fermionic states that are now recognized to be states of spacetime fermions. The resulting, truncated sector is called the R$-$ sector. The R+ sector is defined as the set of R states with $(-1)^F = +1$. At each mass level it contains the same number of states as the R$-$ sector.

After this truncation, the generating function (14.68) for the Ramond sector reduces to

$$f_{\mathrm{R}-}(x) = 8 \prod_{n=1}^{\infty} \Big(\frac{1+x^n}{1-x^n} \Big)^8 , \tag{14.70}$$

since each combination of oscillators now acts on eight ground states only, the ones of the type required to get $(-1)^F = -1$. Expanding the above power series we have

$$f_{\mathrm{R}-}(x) = 8 + 128\,x + 1152\,x^2 + 7680\,x^3 + 42112\,x^4 + \cdots. \tag{14.71}$$

There are eight massless fermionic states.

Let us now reconsider the states of the NS sector. No states here carry spinor indices, so we attempt to get spacetime bosons from this sector. The ground states $|\mathrm{NS}\rangle \otimes |p^+, \vec{p}_T\rangle$ are tachyonic and have $(-1)^F = -1$. The massless states $b^I_{-1/2}|\mathrm{NS}\rangle \otimes |p^+, \vec{p}_T\rangle$ carry a Lorentz vector index, so they are naturally identified as the eight photon states that arise from a ten-dimensional Maxwell gauge field. We wish to keep these eight states to match the eight massless fermionic states of the R sector. Since all bosons should arise from states with the same value of $(-1)^F$, we truncate the NS sector to the set of states with $(-1)^F = +1$. The resulting states comprise the so-called NS+ sector. The NS+ sector contains the massless states and throws away the tachyonic states. Moreover, as explained earlier, the states with $(-1)^F = +1$ all have an odd number of fermionic oscillators and integer $\alpha' M^2$. The set of mass-squared levels in the NS+ sector coincides with the set of mass-squared levels in the R− sector. The NS− sector is defined to contain all the states of the NS sector with $(-1)^F = -1$. The NS− sector contains a tachyon.

Our results strongly hint that the full open string theory, defined by combining additively the set of states from the R− and NS+ sectors, has a supersymmetric spectrum. Indeed, the integer mass-squared levels in the NS generating function (14.67) have degeneracies that match those of (14.71) for the R− sector.

In order to see if the number of fermionic and bosonic states match for all levels we need a generating function $f_{\mathrm{NS}+}(x)$ for the NS+ sector. If we take $f_{\mathrm{NS}}(x)$ in (14.66) and change the sign inside each factor in the numerator

$$\frac{1}{\sqrt{x}} \prod_{n=1}^{\infty} \Big(\frac{1 - x^{n-\frac{1}{2}}}{1 - x^n} \Big)^8, \tag{14.72}$$

the *only* effect is changing the sign of each term in the generating function whose states arise with an odd number of fermions. Since these are precisely the states we want to keep, we can obtain the desired generating function by subtracting (14.72) from (14.66) and dividing by two

$$f_{\mathrm{NS}+}(x) = \frac{1}{2\sqrt{x}} \Big[\prod_{n=1}^{\infty} \Big(\frac{1 + x^{n-\frac{1}{2}}}{1 - x^n} \Big)^8 - \prod_{n=1}^{\infty} \Big(\frac{1 - x^{n-\frac{1}{2}}}{1 - x^n} \Big)^8 \Big]. \tag{14.73}$$

In order to have spacetime supersymmetry we need $f_{\mathrm{NS}+}(x) = f_{\mathrm{R}-}(x)$, or explicitly:

$$\frac{1}{2\sqrt{x}} \Big[\prod_{n=1}^{\infty} \Big(\frac{1 + x^{n-\frac{1}{2}}}{1 - x^n} \Big)^8 - \prod_{n=1}^{\infty} \Big(\frac{1 - x^{n-\frac{1}{2}}}{1 - x^n} \Big)^8 \Big] = 8 \prod_{n=1}^{\infty} \Big(\frac{1 + x^n}{1 - x^n} \Big)^8. \tag{14.74}$$

This intricate identity was proven by Carl Gustav Jacob Jacobi in his treatise on elliptic functions, published in 1829. Jacobi called (14.74) an obscure identity: "*aequatio identica satis abstrusa*". Presently, we recognize it as a key equation at the basis of supersymmetric

string theory. The critical dimension of ten is also visible from the powers of eight on both sides of the equation.

The constructed theory of open superstrings is the theory of a single space-filling stable D9-brane. The D-brane is stable since the theory has no tachyon.

14.8 Closed string theories

We saw in Chapter 13 that closed strings are roughly obtained by combining multiplicatively left-moving and right-moving copies of an open string theory. The same is true for closed superstring theory. Since an open superstring has two sectors (NS and R), closed string sectors can be formed in four ways by combining a left-moving sector (NS or R) with a right-moving sector (NS or R). The result is four closed string sectors:

$$\text{closed string sectors:} \quad (\text{NS}, \text{NS}), \quad (\text{NS}, \text{R}), \quad (\text{R}, \text{NS}), \quad (\text{R},\text{R})\ . \tag{14.75}$$

Conventionally, the first input in $(\cdot, \cdot)$ is the left-moving sector and the second input is the right-moving sector. We also have operators $(-1)^{F_L}$ and $(-1)^{F_R}$ that count fermions in the L and R sectors, respectively. In open superstrings spacetime bosons arise from the NS sector and spacetime fermions arise from the R sector. In closed superstring theories spacetime bosons arise from the (NS,NS) sector and also from the (R,R) sector, since this sector is "doubly" fermionic. The spacetime fermions arise from the (NS,R) and (R, NS) sectors.

In order to get a closed string theory with supersymmetry we must truncate the four sectors above. A consistent truncation arises if we use truncated left and right sectors to begin with. Suppose we take, for example,

$$\text{left sector:} \begin{Bmatrix} \text{NS}+ \\ \text{R}- \end{Bmatrix}, \quad \text{right sector:} \begin{Bmatrix} \text{NS}+ \\ \text{R}+ \end{Bmatrix}. \tag{14.76}$$

Combining these sectors multiplicatively we find the four sectors of the type IIA superstring:

$$\text{type IIA:} \ \ (\text{NS}+, \text{NS}+), \ \ (\text{NS}+, \text{R}+), \ \ (\text{R}-, \text{NS}+), \ \ (\text{R}-, \text{R}+). \tag{14.77}$$

In a closed string theory the value of the mass-squared is given by

$$\tfrac{1}{2}\alpha' M^2 = \alpha' M_L^2 + \alpha' M_R^2, \tag{14.78}$$

where M_L^2 and M_R^2 denote the mass-squared operators for the open string theories that are used to build the left and right sectors, respectively. As befits closed strings there is also the level-matching condition $\alpha_0^- = \bar{\alpha}_0^-$ on the states. This condition guarantees that the left and right sectors give identical contributions to the mass-squared: $\alpha' M_L^2 = \alpha' M_R^2$. No closed string states can be formed if the left and right mass levels do not match.

The type IIA superstring has no tachyons and its massless states are obtained by combining the massless states of the various sectors:

$$(\text{NS+, NS+}): \quad \bar{b}^I_{-1/2}|\text{NS}\rangle_L \otimes b^J_{-1/2}|\text{NS}\rangle_R \otimes |p^+, \vec{p}_T\rangle, \tag{14.79}$$

$$(\text{NS+, R+}): \quad \bar{b}^I_{-1/2}|\text{NS}\rangle_L \otimes |\text{R}_{\bar{b}}\rangle_R \otimes |p^+, \vec{p}_T\rangle, \tag{14.80}$$

$$(\text{R}-, \text{NS+}): \quad |\text{R}_a\rangle_L \otimes b^I_{-1/2}|\text{NS}\rangle_R \otimes |p^+, \vec{p}_T\rangle, \tag{14.81}$$

$$(\text{R}-, \text{R+}): \quad |\text{R}_a\rangle_L \otimes |\text{R}_{\bar{b}}\rangle_R \otimes |p^+, \vec{p}_T\rangle. \tag{14.82}$$

The states in (14.79) carry two independent vector indices I, J that run over eight values. There are therefore 64 bosonic states. Just like the massless states in bosonic closed string theory they carry two vector indices. We therefore get a graviton, a Kalb–Ramond field, and a dilaton:

$$(\text{NS+, NS+}) \text{ massless fields}: \; g_{\mu\nu}, \; B_{\mu\nu}, \; \phi. \tag{14.83}$$

Quick calculation 14.7 Count the number of graviton, Kalb–Ramond, and dilaton states in ten dimensions. Add these numbers up and confirm that you get 64.

The states in (14.80) and (14.81) include only one Ramond vacuum and are therefore spacetime fermions. With $a = 1, \ldots, 8$ and $\bar{b} = \bar{1}, \ldots, \bar{8}$, these give a total of $2 \times 8 \times 8 = 128$ fermionic states. Finally, the states in (14.82) include the product of two R ground states, they are "doubly" fermionic, and thus spacetime *bosons*. There are $8 \times 8 = 64$ massless (R−, R+) bosonic states. Together with the NS–NS states in (14.79), they add up to the 128 massless bosonic states of the closed superstring. As required by supersymmetry, these match with the 128 massless fermionic states of the R–NS and NS–R sectors.

It should be said that the same type IIA string theory arises if the R+ and R− sectors in (14.76) where interchanged. Plainly, the type IIA superstring arises when the left and right truncated R sectors are of different types.

A different theory, a type IIB superstring, arises if the chosen Ramond sectors are of the same type:

$$\text{left}: \left\{\begin{matrix}\text{NS+}\\ \text{R}-\end{matrix}\right\}, \quad \text{right}: \left\{\begin{matrix}\text{NS+}\\ \text{R}-\end{matrix}\right\}. \tag{14.84}$$

We then get

$$\text{type IIB}: \; (\text{NS+, NS+}), \; (\text{NS+, R}-), \; (\text{R}-, \text{NS+}), \; (\text{R}-, \text{R}-). \tag{14.85}$$

The massless states of this theory are

$$(\text{NS+, NS+}): \quad \bar{b}^I_{-1/2}|\text{NS}\rangle_L \otimes b^J_{-1/2}|\text{NS}\rangle_R \otimes |p^+, \vec{p}_T\rangle, \tag{14.86}$$

$$(\text{NS+, R}-): \quad \bar{b}^I_{-1/2}|\text{NS}\rangle_L \otimes |\text{R}_b\rangle_R \otimes |p^+, \vec{p}_T\rangle, \tag{14.87}$$

$$(\text{R}-, \text{NS+}): \quad |\text{R}_a\rangle_L \otimes b^I_{-1/2}|\text{NS}\rangle_R \otimes |p^+, \vec{p}_T\rangle, \tag{14.88}$$

$$(\text{R}-, \text{R}-): \quad |\text{R}_a\rangle_L \otimes |\text{R}_b\rangle_R \otimes |p^+, \vec{p}_T\rangle. \tag{14.89}$$

The same type IIB theory would arise had we replaced both R− sectors in (14.84) by R+ sectors.

While the (NS+, NS+) bosons of type IIA and type IIB theories are the same, the R–R bosons are rather different. In the type IIA theory the massless R–R bosons include a Maxwell field A_μ and a three-index antisymmetric gauge field $A_{\mu\nu\rho}$. In the type IIB theory the massless R–R bosons include a scalar field A, a Kalb–Ramond field $A_{\mu\nu}$, and a totally antisymmetric gauge field $A_{\mu\nu\rho\sigma}$ with four indices. Summarizing

$$\text{R–R massless fields in type IIA}: \quad A_\mu, \quad A_{\mu\nu\rho}, \tag{14.90}$$

$$\text{R–R massless fields in type IIB}: \quad A, \quad A_{\mu\nu}, A_{\mu\nu\rho\sigma}. \tag{14.91}$$

The above R–R fields are deeply related to the existence of *stable* D-branes in type II superstring theories. We discuss this in Section 16.4, where we explain that stable D-branes are actually charged. In bosonic string theory all Dp-branes are unstable and none of them is charged.

The two truncations of (14.75) discussed above led to supersymmetric closed string theories. Other truncations of (14.75) are actually consistent, but the resulting theories are not supersymmetric. These truncations use the NS− sector, leading to a spectrum with tachyons.

Quick calculation 14.8 What sector(s) can be combined with a left-moving NS− to form a consistent closed string sector?

In addition to the type II theories, there are also two *heterotic* superstring theories. These are remarkable closed string theories. While a type II closed superstring arises by combining together left-moving and right-moving copies of open superstrings, in the heterotic string we combine a left-moving open *bosonic* string with a right-moving open *superstring*! Out of the 26 left-moving bosonic coordinates of the bosonic factor only ten of them are matched by the right-moving bosonic coordinates of the superstring factor. As a result, this theory effectively lives in ten-dimensional spacetime. Heterotic strings come in two versions: $E_8 \times E_8$ type and $SO(32)$ type. These labels characterize the groups of symmetries that exist in the theories. E_8 is a group, in fact, it is the largest exceptional group (the E is for exceptional). The group $SO(32)$ is the group generated by 32-by-32 matrices that are orthogonal and have unit determinant. A discussion of the heterotic $SO(32)$ theory can be found in Problem 14.5.

Finally, in addition to both type II and heterotic theories, there is the type I theory. This is a supersymmetric theory of open and closed *unoriented* strings. A string theory is unoriented (see Problems 12.12 and 13.5) if the states of the theory are invariant under an operation that reverses the orientation of the strings. Both the type II theories and the heterotic theories are theories of oriented closed strings.

The complete list of ten-dimensional supersymmetric string theories is therefore

- type IIA,
- type IIB,
- $E_8 \times E_8$ heterotic,

- $SO(32)$ heterotic,
- type I.

These five theories have all been known since the middle of the 1980s. Some relationships between them were found soon after their discovery, but a clearer picture emerged only in the late 1990s. The limit of type IIA theory as the string coupling is taken to infinity was shown to give a theory in eleven dimensions. This theory is called M-theory, with the meaning of M to be decided when the nature of the theory becomes clear. It is known, however, that M-theory is *not* a string theory. M-theory contains membranes (2-branes) and 5-branes, and these branes are not D-branes. M-theory may end up playing a prominent role in understanding string theory. The discovery of many other relationships between the five string theories listed above and M-theory has made it clear that we really have just *one theory*. This is a fundamental result: there is a unique theory, and the five superstrings and M-theory are different limits of this unique theory.

It is not clear if bosonic strings are part of this interrelated set of theories. They certainly seem quite different from the superstrings. It would be very interesting, however, if *all* string theories were one single theory. There have been suggestions that bosonic string theories and superstrings are related via cosmological solutions. We have certainly not heard the last word on this subject.

Problems

Problem 14.1 Counting bosonic states.

(a) Consider k ordinary commuting oscillators a^i, with $i = 1, \ldots, k$. How many products of the form $a^{i_1} a^{i_2}$ can be built? How many $a^{i_1} a^{i_2} a^{i_3}$? How many $a^{i_1} a^{i_2} a^{i_3} a^{i_4}$? [Hint: use the result in Problem 12.11.]

(b) List and count the states in the $\alpha' M^2 = 3$ level of the open bosonic string. Confirm that you get the same number of states predicted by the generating function $f_{os}(x)$ in (14.63).

Problem 14.2 Generating function for the unoriented bosonic open string theory.

Write a generating function for the unoriented bosonic open string theory by starting with the generating function $f_{os}(x)$ for the full oriented theory and adding a term that implements the projection to unoriented states.

Problem 14.3 Massive level in the open superstring.

(a) Consider eight anticommuting variables b^i, with $i = 1, \ldots, 8$. Ignoring signs, how many inequivalent products of the form $b^{i_1} b^{i_2}$ can be built? How many $b^{i_1} b^{i_2} b^{i_3}$? How many $b^{i_1} b^{i_2} b^{i_3} b^{i_4}$?

(b) Consider the first and second excited levels of the open superstring ($\alpha' M^2 = 1$ and $\alpha' M^2 = 2$). List the states in the NS sector and the states in the R sector. Confirm that you get the same number of states.

Problem 14.4 Closed string degeneracies.

For closed string states the left-moving and right-moving excitations are each described like states of open strings with identical values of $\alpha' M^2$. The value of $\alpha' M^2$ for the closed string state is four times that value.

(a) State the values of $\alpha' M^2$ and give the degeneracies for the first five mass levels of the closed bosonic string theory.
(b) State the values of $\alpha' M^2$ and give the separate degeneracies of bosons and fermions for the first five mass levels of the type IIA closed superstrings. Would the answer have been different for type IIB?

Problem 14.5 Counting states in heterotic $SO(32)$ string theory.

In heterotic (closed) string theory the right-moving part of the theory is that of an open superstring. It has an NS sector whose states are built with oscillators α^I_{-n} and b^I_{-r} acting on the NS vacuum. It also has an R sector whose states are built with oscillators α^I_{-n} and d^I_{-n} acting on the R ground states. The index I runs over 8 values. The standard GSO projection down to NS+ and R− applies.

The left-moving part of the theory is that of a peculiar bosonic open string. The 24 transverse coordinates split into eight bosonic coordinates X^I with oscillators $\bar{\alpha}^I_{-n}$ and 16 peculiar bosonic coordinates. A surprising fact of two-dimensional physics allows us to replace these 16 coordinates by 32 two-dimensional left-moving *fermion* fields λ^A, with $A = 1, 2, \ldots, 32$. The (anticommuting) fermion fields λ^A imply that the left-moving part of the theory also has NS′ and R′ sectors, denoted with primes to differentiate them from the standard NS and R sectors of the open superstring.

The left NS′ sector is built with oscillators $\bar{\alpha}^I_{-n}$ and λ^A_{-r} acting on the vacuum $|\text{NS}'\rangle_L$, declared to have $(-1)^{F_L} = +1$:

$$(-1)^{F_L}|\text{NS}'\rangle_L = +|\text{NS}'\rangle_L.$$

The *naive* mass formula in this sector is

$$\alpha' M_L^2 = \frac{1}{2}\sum_{n\neq 0} \bar{\alpha}^I_{-n}\bar{\alpha}^I_n + \frac{1}{2}\sum_{r\in\mathbb{Z}+\frac{1}{2}} r\,\lambda^A_{-r}\lambda^A_r.$$

The left R′ sector is built with oscillators $\bar{\alpha}^I_{-n}$ and λ^A_{-n} acting on a set of R′ ground states. The naive mass formula in this sector is

$$\alpha' M_L^2 = \frac{1}{2}\sum_{n\neq 0}\left(\bar{\alpha}^I_{-n}\bar{\alpha}^I_n + n\,\lambda^A_{-n}\lambda^A_n\right).$$

Momentum labels are not needed in this problem so they are omitted throughout.

(a) Consider the left NS′ sector. Write the precise mass-squared formula with normal-ordered oscillators and the appropriate normal-ordering constant. The GSO projection here keeps the states with $(-1)^{F_L} = +1$; this defines the left NS′+ sector. Write

explicitly and count the states we keep for the three lowest mass levels, indicating the corresponding values of $\alpha' M_L^2$. [This is a long list.]

(b) Consider the left R′ sector. Write the precise mass-squared formula with normal-ordered oscillators and the appropriate normal-ordering constant. We have 32 zero modes λ_0^A and 16 linear combinations behave as creation operators. As usual, half of the ground states have $(-1)^{F_L} = +1$ and the other half have $(-1)^{F_L} = -1$. Let $|R_\alpha\rangle_L$ denote ground states with $(-1)^{F_L} = +1$. How many ground states $|R_\alpha\rangle_L$ are there? Keep only states with $(-1)^{F_L} = +1$; this defines the left R′+ sector. Write explicitly and count the states we keep for the two lowest mass levels, indicating the corresponding values of $\alpha' M_L^2$. [This is a shorter list.]

At any mass level $\alpha' M^2 = 4k$ of the *heterotic* string, the spacetime bosons are obtained by "tensoring" *all* the left states (NS′+ and R′+) with $\alpha' M_L^2 = k$ with the right-moving NS+ states with $\alpha' M_R^2 = k$. Similarly, the spacetime fermions are obtained by tensoring all the left states (NS′+ and R′+) with $\alpha' M_L^2 = k$ with the right-moving R− states with $\alpha' M_R^2 = k$. At any mass level where either left states or right states are missing, one cannot form heterotic string states.

(c) Are there tachyonic states in heterotic string theory? Write out the massless states of the theory (bosons and fermions) and describe the fields associated with the bosons. Calculate the total number of states in heterotic string theory (bosons plus fermions) at $\alpha' M^2 = 4$. *Answer:* 18 883 584 states.

(d) Write a generating function $f_L(x) = \sum_r a(r) x^r$ for the full set of GSO-truncated states in the left-moving sector (include both NS′+ and R′+ states). Use the convention where $a(r)$ counts the number of states with $\alpha' M_L^2 = r$. Use $f_L(x)$ and an algebraic manipulator to find the total number of states in heterotic string theory at $\alpha' M^2 = 8$. *Answer*: 6 209 372 160.

PART II

DEVELOPMENTS

15 D-branes and gauge fields

The open strings we have studied so far were described by coordinates all of which satisfy Neumann boundary conditions. These open strings move on the world-volume of a space-filling D25-brane. Here we quantize open strings attached to more general D-branes. We begin with the case of a single Dp-brane, with $1 \leq p < 25$. We then turn to the case of multiple parallel Dp-branes, where we see the appearance of interacting gauge fields and the possibility of massive gauge fields. We continue with the case of parallel D-branes of different dimensionalities.

15.1 Dp-branes and boundary conditions

A Dp-brane is an extended object with p *spatial* dimensions. In bosonic string theory, where the number of spatial dimensions is 25, a D25-brane is a space-filling brane. The letter D in Dp-brane stands for Dirichlet. In the presence of a D-brane, the endpoints of open strings must lie on the brane. As we will see in more detail below, this requirement imposes a number of Dirichlet boundary conditions on the motion of the open string endpoints.

Not all extended objects in string theory are D-branes. Strings, for example, are 1-branes because they are extended objects with one spatial dimension, but they are not D1-branes. Branes with p spatial dimensions are generically called p-branes. A 0-brane is some kind of particle. Just as the world-line of a particle is one-dimensional, the world-volume of a p-brane is $(p+1)$-dimensional. Of these $p+1$ dimensions, one is the time dimension and the other p are spatial dimensions. We first discussed the concept of D-branes in Section 6.5. In addition, Problem 6.11 examined the classical motion of open strings ending on D-branes of various dimensionalities. Our main subject in the present chapter is the quantization of open strings in the presence of various kinds of D-branes. This is a rich subject with important implications for the problem of constructing realistic physical models using strings. Furthermore, the study of D-branes and the gravitational fields they produce has led to surprising new insights in the study of strongly interacting gauge theories.

In this section, we set up the notation needed to describe D-branes, and then we state the appropriate boundary conditions. We let d denote the total number of spatial dimensions in the theory; in the present case, $d = 25$. The total number of spacetime dimensions is $D = d + 1 = 26$. A Dp-brane with $p < 25$ extends over a p-dimensional subspace of the 25-dimensional space. We will focus on simple Dp-branes: those that are p-dimensional

hyperplanes inside the d-dimensional space. How can we specify such hyperplanes? We need $(d-p)$ linear conditions. In three spatial dimensions ($d=3$), a 2-brane ($p=2$) is a plane, and it is specified by one linear condition ($d-p=3-2=1$). For example, $z=0$, specifies the (x, y) plane. Similarly, a string along the z axis ($p=1$) is specified by two linear conditions ($d-p=3-1=2$): $x=0$ and $y=0$. We need as many conditions as there are spatial coordinates normal to the brane.

Consider now a Dp-brane. We introduce spacetime coordinates x^μ, with $\mu=0, 1, \ldots, 25$, that are split into two groups. The first group includes the coordinates tangential to the brane *world-volume*. These are the time coordinate and p spatial coordinates. The second group includes the $(d-p)$ coordinates normal to the brane world-volume. We write

$$\underbrace{x^0,\, x^1,\, \ldots,\, x^p}_{\text{D}p \text{ tangential coordinates}} \quad \underbrace{x^{p+1},\, x^{p+2},\, \ldots,\, x^d}_{\text{D}p \text{ normal coordinates}}. \tag{15.1}$$

The location of the Dp-brane is specified by fixing the values of the coordinates normal to the brane. With this split in mind we write

$$x^a = \bar{x}^a\,, \quad a = p+1, \ldots, d. \tag{15.2}$$

Here the $\bar{x}^a$ are a set of $(d-p)$ constants. In a completely analogous fashion, the string coordinates $X^\mu(\tau,\sigma)$ are split as

$$\underbrace{X^0,\, X^1,\, \ldots,\, X^p}_{\text{D}p \text{ tangential coordinates}} \quad \underbrace{X^{p+1},\, X^{p+2},\, \ldots,\, X^d}_{\text{D}p \text{ normal coordinates}}. \tag{15.3}$$

Since the endpoints of the open string must lie on the Dp-brane, the string coordinates normal to the brane must satisfy Dirichlet boundary conditions

$$X^a(\tau,\sigma)\Big|_{\sigma=0} = X^a(\tau,\sigma)\Big|_{\sigma=\pi} = \bar{x}^a\,, \quad a = p+1, \ldots, d. \tag{15.4}$$

The string coordinates X^a are called DD coordinates, because both endpoints satisfy a Dirichlet boundary condition. The open string endpoints can move freely along the directions tangential to the D-brane. As a result, the string coordinates tangential to the D-brane satisfy Neumann boundary conditions:

$$X^{m\prime}(\tau,\sigma)\Big|_{\sigma=0} = X^{m\prime}(\tau,\sigma)\Big|_{\sigma=\pi} = 0\,, \quad m = 0, 1, \ldots, p. \tag{15.5}$$

These string coordinates are called NN coordinates because both endpoints satisfy a Neumann boundary condition. We see that the split (15.3) into tangential and normal coordinates is also a split into coordinates which satisfy Neumann and Dirichlet boundary conditions, respectively:

$$\underbrace{X^0,\, X^1,\, \ldots,\, X^p}_{\text{NN coordinates}} \quad \underbrace{X^{p+1},\, X^{p+2},\, \ldots,\, X^d}_{\text{DD coordinates}}. \tag{15.6}$$

In order to use the light-cone gauge we need at least one spatial NN coordinate that can be used together with X^0 to define the coordinates $X^\pm$. We therefore need to assume $p \geq 1$, and our analysis does not apply to strings attached to a D0-brane. D0-branes are perfectly

consistent objects, but in order to study them we need a Lorentz covariant quantization (see Chapter 24 and Problem 24.4). We will label the light-cone coordinates as

$$\underbrace{X^+, X^-, \{X^i\}}_{\text{NN}} \quad \underbrace{\{X^a\}}_{\text{DD}} \qquad i = 2, \ldots, p \quad \text{and} \quad a = p+1, \ldots, d. \tag{15.7}$$

15.2 Quantizing open strings on Dp-branes

Having specified the boundary conditions on the various string coordinates, we can proceed to the quantization of open strings in the presence of a Dp-brane. The purpose of the analysis that follows is to determine the spectrum of open string states and to use this result to understand more deeply what goes on in the world-volume of a Dp-brane.

Our earlier work in Chapter 12 is quite useful here. The NN coordinates $X^i(\tau, \sigma)$ satisfy exactly the same conditions that are satisfied by the light-cone coordinates $X^I(\tau, \sigma)$ of open strings attached to a D25-brane. All expansions and commutation relations for the X^i coordinates can be obtained from those of X^I by replacing $I \to i$ in the relevant equations.

We recall that the X^- coordinate was determined in terms of the transverse light-cone coordinates in equation (9.65):

$$\dot{X}^- \pm X^{-\prime} = \frac{1}{2\alpha'} \frac{1}{2p^+} (\dot{X}^I \pm X^{I\prime})^2. \tag{15.8}$$

Moreover, the mode expansion of $\dot{X}^I \pm X^{I\prime}$ was given in (9.74):

$$\dot{X}^I \pm X^{I\prime} = \sqrt{2\alpha'} \sum_{n \in \mathbb{Z}} \alpha_n^I \, e^{-in(\tau \pm \sigma)}. \tag{15.9}$$

A completely analogous mode expansion held for the coordinate X^-; this expansion continues to hold without change since X^- remains an NN coordinate. The above equations, together with the X^- expansion, led to (12.105) and (12.106), summarized here as

$$2p^+ p^- \equiv \frac{1}{\alpha'} \Big(\frac{1}{2} \alpha_0^I \, \alpha_0^I + \sum_{n=1}^{\infty} \alpha_{-n}^I \, \alpha_n^I + a \Big). \tag{15.10}$$

The ordering constant a was determined to be equal to minus one for the quantization of strings on a D25-brane. The light-cone index $I = 2, \ldots, 25$, takes values that, for a Dp-brane, run over NN coordinates labeled by i and DD coordinates labeled by a. As a result, (15.8) now becomes

$$\dot{X}^- \pm X^{-\prime} = \frac{1}{2\alpha'} \frac{1}{2p^+} \Big\{ (\dot{X}^i \pm X^{i\prime})^2 + (\dot{X}^a \pm X^{a\prime})^2 \Big\}. \tag{15.11}$$

As explained before, the X^i coordinates are expanded as

$$\dot{X}^i \pm X^{i\prime} = \sqrt{2\alpha'} \sum_{n \in \mathbb{Z}} \alpha_n^i \, e^{-in(\tau \pm \sigma)}. \tag{15.12}$$

The X^a coordinates are the ones we must investigate. If an expansion analogous to (15.12) holds for X^a, we will be able to find p^- by letting $I \to (i, a)$ in (15.10), just as we did in order to obtain (15.11).

We are now in a position to address the novel part of the quantization of open strings attached to a Dp-brane. The coordinates X^a normal to the brane satisfy the wave equation, so the general solution is a superposition of two waves:

$$X^a(\tau, \sigma) = \frac{1}{2}\Big(f^a(\tau + \sigma) + g^a(\tau - \sigma)\Big). \tag{15.13}$$

Let us examine the boundary conditions (15.4). At $\sigma = 0$ we obtain

$$X^a(\tau, 0) = \frac{1}{2}\big(f^a(\tau) + g^a(\tau)\big) = \bar{x}^a, \tag{15.14}$$

so that $g^a(\tau) = -f^a(\tau) + 2\bar{x}^a$, and as a result,

$$X^a(\tau, \sigma) = \bar{x}^a + \frac{1}{2}\Big(f^a(\tau + \sigma) - f^a(\tau - \sigma)\Big). \tag{15.15}$$

The boundary condition at $\sigma = \pi$ then gives us

$$f^a(\tau + \pi) = f^a(\tau - \pi). \tag{15.16}$$

This simply means that $f^a(u)$ is a periodic function with period 2π. This information is incorporated into the following expansion:

$$f^a(u) = \tilde{f}^a_0 + \sum_{n=1}^{\infty}\Big(\tilde{f}^a_n \cos nu + \tilde{g}^a_n \sin nu\Big). \tag{15.17}$$

It is interesting to note that there is no term linear in u. Such a term was present when the coordinate satisfied a Neumann boundary condition because in that case it was the derivative $f'(u)$ which was periodic. Replacing (15.17) in (15.15) and performing some trigonometric simplification, we find

$$X^a(\tau, \sigma) = \bar{x}^a + \sum_{n=1}^{\infty}\Big(-\tilde{f}^a_n \sin n\tau \sin n\sigma + \tilde{g}^a_n \cos n\tau \sin n\sigma\Big). \tag{15.18}$$

Redefining the expansion coefficients that are arbitrary anyway, we can write

$$X^a(\tau, \sigma) = \bar{x}^a + \sum_{n=1}^{\infty}\Big(f^a_n \cos n\tau + \tilde{f}^a_n \sin n\tau\Big) \sin n\sigma. \tag{15.19}$$

Since there is no term linear in τ, the string has no net time-averaged momentum in the x^a direction. This is reasonable since strings must remain attached to the brane. If there were a $p^a\tau$ term present, the endpoints $\sigma = 0, \pi$ would not remain at $x^a = \bar{x}^a$ when $\tau \neq 0$.

In order to define the quantum theory associated with X^a, we focus on the classical parameters that describe the motion of the open string in equation (15.19). Since we are

trying to quantize strings attached to a *fixed* Dp-brane, the values $\bar{x}^a$ are not parameters that can be adjusted to describe various open string motions. The $(f^a, \tilde{f}^a)$, on the other hand, *are* parameters of the open string motion. Therefore, in quantizing the open string, the $\bar{x}^a$ remain numbers and *do not* become operators, while the $(f^a, \tilde{f}^a)$ turn into operators.

We now rewrite (15.19) in terms of oscillators, defined in order to simplify the following analysis:

$$X^a(\tau,\sigma) = \bar{x}^a + \sqrt{2\alpha'}\sum_{n\neq 0}\frac{1}{n}\alpha_n^a e^{-in\tau}\sin n\sigma. \tag{15.20}$$

The string coordinate X^a is Hermitian if $(\alpha_n^a)^\dagger = \alpha_{-n}^a$, which is the usual Hermiticity property of oscillators. Note that the zero mode α_0^a does not exist. Additionally,

$$\dot{X}^a = -i\sqrt{2\alpha'}\sum_{n\neq 0}\alpha_n^a e^{-in\tau}\sin n\sigma\,, \quad X^{a\prime} = \sqrt{2\alpha'}\sum_{n\neq 0}\alpha_n^a e^{-in\tau}\cos n\sigma, \tag{15.21}$$

and therefore

$$X^{a\prime} \pm \dot{X}^a = \sqrt{2\alpha'}\sum_{n\neq 0}\alpha_n^a e^{-in(\tau\pm\sigma)}. \tag{15.22}$$

The analogy with (15.12) is quite close, but there are two differences. First, when the lower sign applies, the combinations of derivatives differ by an overall minus sign. Second, the zero mode is absent in (15.22).

The quantization is now straightforward. With $\mathcal{P}^{\tau\,a}(\tau,\sigma) = \dot{X}^a/2\pi\alpha'$, the nonvanishing commutators are postulated to be

$$\Big[\, X^a(\tau,\sigma)\,,\, \dot{X}^b(\tau,\sigma')\,\Big] = 2\pi\alpha'\, i\,\delta^{ab}\,\delta(\sigma-\sigma'). \tag{15.23}$$

Following the analysis of Section 12.2, this commutator can be rewritten in the form (12.30), with (I, J) replaced by (a, b). Since the mode expansions (15.22) take the standard form, the earlier analysis applies. The overall sign difference alluded to above is of no import since $(X^{a\prime} - \dot{X}^a)$ appears twice in the relevant commutators. We thus find

$$[\alpha_m^a, \alpha_n^b] = m\,\delta^{ab}\,\delta_{m+n,0}\,, \quad m, n \neq 0. \tag{15.24}$$

The zero modes work out consistently: $\bar{x}^a$ is a constant, and there is no conjugate momentum since $\alpha_0^a \equiv 0$. The sign difference is also immaterial for the evaluation of (15.11) since $(X^{a\prime} - \dot{X}^a)$ appears squared. Therefore, equation (15.10) can be split as

$$2p^+\,p^- \equiv \frac{1}{\alpha'}\Big(\alpha'\,p^i\,p^i + \sum_{n=1}^{\infty}\Big[\alpha_{-n}^i\,\alpha_n^i + \alpha_{-n}^a\,\alpha_n^a\Big] - 1\Big). \tag{15.25}$$

A few comments are needed here. Since $p^a \sim \alpha_0^a \equiv 0$, the term $\frac{1}{2}\alpha_0^I\alpha_0^I$ simply became $\alpha' p^i p^i$ (recall that $\alpha_0^\mu = \sqrt{2\alpha'}p^\mu$). The ordering constant has been set to minus one, as for the D25-brane. The critical dimension has not been changed either. This is reasonable since

only the zero mode structure is different in X^a and X^i. In particular, the naive contributions needed to normal order $L_0^\perp$ are the same for X^a and for X^i. It follows from (15.25) that

$$M^2 = -p^2 = 2p^+p^- - p^i p^i = \frac{1}{\alpha'}\Big(\sum_{n=1}^{\infty}\Big[\alpha^i_{-n}\alpha^i_n + \alpha^a_{-n}\alpha^a_n\Big] - 1\Big). \tag{15.26}$$

Using creation and annihilation operators, we get

$$M^2 = \frac{1}{\alpha'}\Big(-1 + \sum_{n=1}^{\infty}\sum_{i=2}^{p} n\, a_n^{i\,\dagger} a_n^i + \sum_{m=1}^{\infty}\sum_{a=p+1}^{d} m\, a_m^{a\,\dagger} a_m^a\Big). \tag{15.27}$$

Let us now consider the state space of the quantum string. The ground states in the D25-brane background were $|p^+, \vec{p}_T\rangle$, where $\vec{p}_T = (p^2, \ldots, p^{25})$ is the vector with components p^I. The I index now runs over both i and a values, but there are no p^a operators, so the ground states of the theory are labeled by p^+ and p^i only:

$$|p^+, \vec{p}\,\rangle \quad \text{with} \quad \vec{p} = (p^2, \ldots, p^p). \tag{15.28}$$

We build additional states by acting with oscillators on the ground states. We have oscillators along the brane:

$$a_n^{i\,\dagger}, n \geq 1, \quad i = 2, \ldots, p, \tag{15.29}$$

and oscillators normal to the brane:

$$a_n^{a\,\dagger}, n \geq 1, \quad a = p+1, \ldots, d. \tag{15.30}$$

So the states take the form

$$\Big[\prod_{n=1}^{\infty}\prod_{i=2}^{p}\Big(a_n^{i\,\dagger}\Big)^{\lambda_{n,i}}\Big]\Big[\prod_{m=1}^{\infty}\prod_{a=p+1}^{d}\Big(a_m^{a\,\dagger}\Big)^{\lambda_{m,a}}\Big]|p^+, \vec{p}\rangle. \tag{15.31}$$

Schrödinger wavefunctions take the schematic form

$$\psi_{i_1\ldots i_p\, a_1\ldots a_q}(\tau, p^+, \vec{p}\,). \tag{15.32}$$

Just like the indices on the oscillators, the indices on the wavefunctions are of two types: indices along the directions tangent to the brane (i-type) and indices along the directions normal to the brane (a-type).

In the field theories that describe the states of the string, the fields take the same form as the string Schrödinger wavefunctions. We can therefore ask: where do the fields corresponding to (15.32) live? Are these fields defined over all of spacetime, or only in some subspace of spacetime?

Since the fields are functions of the momenta p^i, by Fourier transformation they can be viewed as fields that depend on the coordinates x^i. The τ dependence is in fact an x^+ dependence, and the p^+ dependence can be Fourier transformed into an x^- dependence. All together, we have fields that depend on x^+, x^-, and x^i, with $i = 2, \ldots, p$. These are

precisely the $(p+1)$ coordinates that span the world-volume of the Dp-brane. It is reasonable to conclude that the fields actually live on the Dp-brane. Indeed, this world-volume is the only natural candidate for a $(p+1)$-dimensional subspace of spacetime.

Our analysis suggests, but does not prove, that the fields live on the Dp-brane; the positions $\bar{x}^a$ of the Dp-brane did not appear in the states nor in the wavefunctions. How could we *prove* that the fields live on the Dp-brane? We would have to study interactions. Since closed strings have no endpoints, they are not fixed by D-branes and can exist over all of spacetime. By scattering closed strings off the Dp-brane we can investigate whether the interactions between fields from the closed string sector and fields from the open string sector take place on the D-brane world-volume. The answer appears to be yes. Statements about where open string fields live, however, are likely to be ambiguous or even gauge dependent. Different answers could be completely consistent.

We conclude our analysis of the Dp-brane by giving a list and detailed description of the fields that satisfy $M^2 \leq 0$. All these fields live on the Dp-brane, so we must state how they transform under the Lorentz transformations that preserve the Dp-brane. These are Lorentz transformations in $p+1$ dimensions, and the fields may be scalars or vectors, for example. Let us begin with the simplest states, the ground states:

$$|p^+, \vec{p}\,\rangle\,, \quad M^2 = -\frac{1}{\alpha}. \tag{15.33}$$

These states are tachyon states on the brane, and they have exactly the same mass as the tachyon states we found on the D25-brane. The corresponding tachyon field, of course, is just a Lorentz scalar on the brane.

The next states have one oscillator acting on them. Consider first the case in which the oscillator arises from a coordinate tangent to the brane:

$$a_1^{i\,\dagger}|p^+, \vec{p}\,\rangle\,, \quad i = 2, \ldots, p, \quad M^2 = 0. \tag{15.34}$$

For any momenta, these are $(p+1)-2$ massless states. Moreover, the index they carry lives on the brane. They are therefore states that transform as a Lorentz vector on the brane. Since the number of states equals the spacetime dimensionality of the brane minus two, these are clearly photon states. The associated field is a Maxwell gauge field living on the brane. This is a fundamental result:

a Dp-brane has a Maxwell field living on its world-volume. (15.35)

Finally, let us consider the case in which the oscillator acting on the ground state arises from a coordinate normal to the brane:

$$a_1^{a\dagger}\,|p^+, \vec{p}\,\rangle, \quad a = p+1, \ldots, d, \quad M^2 = 0. \tag{15.36}$$

For any momenta, these are $(d-p)$ states living on the brane. Since the index a is not a Lorentz index for the brane, this index is merely a counting label. These states transform

as Lorentz *scalars* on the brane. Therefore, we get a massless scalar field for each direction normal to the Dp-brane:

a Dp-brane has a massless scalar for each normal direction. (15.37)

These massless scalars have a physical interpretation. In Section 12.8 we indicated that open string states represent D-brane excitations. Our Dp-brane and a slightly displaced parallel Dp-brane are actually states of the same energy. The displaced Dp-brane can be viewed as a zero-energy excitation of the original brane. It is also an excitation with zero momentum, since it represents a displacement that is constant over all of the Dp-brane. A zero-momentum excitation with zero energy is a massless excitation because it obeys the energy-momentum relation $E = p$ of a massless particle. The massless scalars identified in (15.36) give rise to these excitations. This interpretation is supported by the fact that we have as many massless fields as there are directions normal to the Dp-brane. Those are the independent directions in which the Dp-brane can be moved. Note that a space-filling D25-brane has no massless scalars on its world-volume, consistent with the fact that it cannot be displaced.

All in all, the massless states on the Dp-brane are $(p-1)$ photon states and $(d-p)$ scalar field states. Apart from the momentum labels which are different, we have the same number of massless states as on the D25-brane. The $(d-1)$ states on the D25-brane are accounted for on the Dp-brane by $(p-1)$ photon states and $(d-p)$ scalar states.

15.3 Open strings between parallel Dp-branes

We will now consider the quantization of open strings that extend between two parallel Dp-branes. In describing such branes we will continue to use the notation of the previous sections. Two parallel branes of the same dimensionality have the same set of longitudinal coordinates and the same set of normal coordinates. Recall that the values $\bar{x}^a$ of the normal coordinates specify the position of a Dp-brane. This time the first Dp-brane is located at $x^a = \bar{x}_1^a$ and the second at $x^a = \bar{x}_2^a$. If we happen to have $\bar{x}_1^a = \bar{x}_2^a$ for all a, the two Dp-branes coincide in space – they are on top of each other. Otherwise, they are separated. In Figure 15.1 we show two parallel, separated D2-branes.

What kinds of open strings does this configuration of parallel Dp-branes support? There are actually four different classes of strings, each of which must be analyzed separately. The first two classes are made up of open strings that begin and end on the same D-brane, either brane one or brane two. These strings we already studied and quantized in the previous section. The other two classes consist of strings that start on one brane and end on the other. These are *stretched strings*. The strings that begin on brane one and end on brane

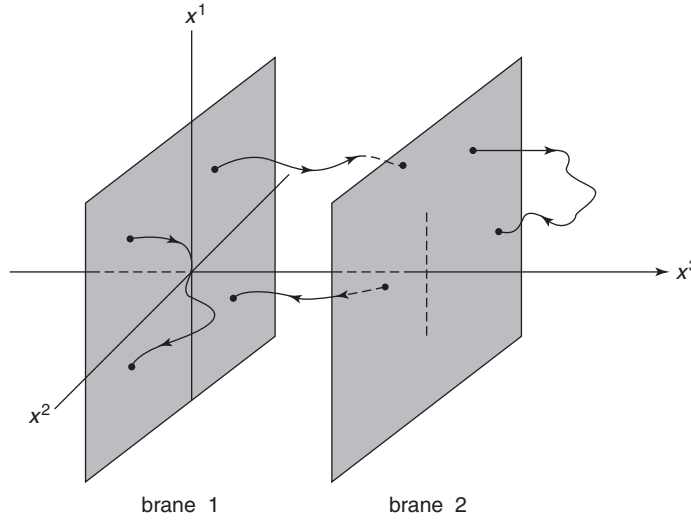

Fig. 15.1 Two parallel D2-branes. Here x^1 and x^2 are longitudinal coordinates, and x^3 is a normal coordinate. The positions of brane one and brane two are specified by the coordinates $\bar{x}_1^3$ and $\bar{x}_2^3$, respectively. We show the four types of strings that this configuration supports.

two are different from the strings that begin on brane two and end on brane one. These strings are oppositely oriented, and the orientation of a string (the direction of increasing σ) matters. As we will show in Chapter 16, the charge of a string changes sign when we reverse its orientation. The classes of open strings that are supported on a particular configuration of D-branes are called *sectors*. The quantum theory of open strings in the presence of two parallel Dp-branes has four sectors. In Figure 15.1 we show a string for each of the four sectors.

Let us consider the sector consisting of open strings that begin on brane one and end on brane two. The NN string coordinates X^+, X^-, and X^i are quantized just as before, since the corresponding boundary conditions are still given by (15.5). On the other hand, for the DD string coordinates the boundary conditions (formerly given by (15.4)) are now

$$X^a(\tau,\sigma)\Big|_{\sigma=0} = \bar{x}_1^a\,, \quad X^a(\tau,\sigma)\Big|_{\sigma=\pi} = \bar{x}_2^a\,, \quad a = p+1,\ldots,d. \tag{15.38}$$

The solution of the wave equation subject to these boundary conditions can be studied starting from (15.15), which already incorporates the boundary condition at $\sigma = 0$. In the present case we just change $\bar{x}^a$ to $\bar{x}_1^a$:

$$X^a(\tau,\sigma) = \bar{x}_1^a + \frac{1}{2}\Big(f^a(\tau+\sigma) - f^a(\tau-\sigma)\Big). \tag{15.39}$$

The boundary condition at $\sigma = \pi$ now gives us

$$f^a(\tau+\pi) - f^a(\tau-\pi) = 2\,(\bar{x}_2^a - \bar{x}_1^a), \tag{15.40}$$

or, equivalently,

$$f^a(u+2\pi) - f^a(u) = 2\,(\bar{x}_2^a - \bar{x}_1^a). \tag{15.41}$$

This means that the derivative $f^{a\prime}(u)$ is a periodic function with period 2π and has an expansion of the type indicated in (15.17). Integrating, the function $f^a(u)$ has an expansion of the form

$$f^a(u) = f_0^a\, u + \sum_{n=1}^{\infty}(h_n^a \cos nu + g_n^a \sin nu). \tag{15.42}$$

We have not included a constant term because it would drop out of X^a, as can be seen in (15.39). The constant f_0^a is fixed by the boundary condition (15.41):

$$f_0^a = \frac{1}{\pi}(\bar{x}_2^a - \bar{x}_1^a). \tag{15.43}$$

It is now possible to substitute $f^a(u)$ into (15.39). Aside from handling the zero modes, the computations are identical to those that led to (15.19). This time we obtain

$$X^a(\tau,\sigma) = \bar{x}_1^a + (\bar{x}_2^a - \bar{x}_1^a)\frac{\sigma}{\pi} + \sum_{n=1}^{\infty}\Big(f_n^a \cos n\tau + \tilde{f}_n^a \sin n\tau\Big)\sin n\sigma. \tag{15.44}$$

Note that the boundary conditions are manifestly satisfied. To describe strings that extend from brane two to brane one we merely exchange $\bar{x}_1^a$ and $\bar{x}_2^a$ in the above equation. We can rewrite (15.44) in terms of oscillators, using (15.20) as a model:

$$X^a(\tau,\sigma) = \bar{x}_1^a + (\bar{x}_2^a - \bar{x}_1^a)\frac{\sigma}{\pi} + \sqrt{2\alpha'}\sum_{n\neq 0}\frac{1}{n}\alpha_n^a\, e^{-in\tau}\sin n\sigma. \tag{15.45}$$

The constants $\bar{x}_1^a$ and $\bar{x}_2^a$ do not become quantum operators because for fixed D-branes, just as before, they are not parameters of the open string fluctuations. Note the absence of terms linear in τ; the open strings have no time-averaged momentum in the x^a directions. Even though we are not giving new names to the oscillators above, they are different operators from those we obtained during the quantization of strings that begin and end on the same Dp-brane. The oscillators in different sectors must not be confused. This time the derivatives give

$$\dot{X}^a = -i\sqrt{2\alpha'}\sum_{n\in\mathbb{Z}}\alpha_n^a e^{-in\tau}\sin n\sigma\,, \qquad X^{a\prime} = \sqrt{2\alpha'}\sum_{n\in\mathbb{Z}}\alpha_n^a e^{-in\tau}\cos n\sigma, \tag{15.46}$$

where

$$\sqrt{2\alpha'}\alpha_0^a = \frac{1}{\pi}\,(\bar{x}_2^a - \bar{x}_1^a). \tag{15.47}$$

Although the strings do not carry momentum in the x^a direction, there is still a nonvanishing α_0^a. There is no contradiction because the interpretation of α_0 as momentum requires that α_0 appear in $\dot{X}$. As you can see, α_0^a appears in $X^{a\prime}$, but not in $\dot{X}^a$. A nonvanishing α_0^a implies stretched strings: α_0^a vanishes precisely when the two D-branes coincide. Similar operators emerge in the expansion of closed strings that wrap around compact dimensions (see Chapter 17).

The two derivatives in (15.46) can be combined to form

$$X^{a\prime} \pm \dot{X}^a = \sqrt{2\alpha'} \sum_{n\in\mathbb{Z}} \alpha_n^a e^{-in(\tau\pm\sigma)}. \tag{15.48}$$

It follows from this result, and our comments in the previous section, that the oscillators satisfy the expected commutation relations. To calculate the mass-squared operator, we reconsider equation (15.10). As before, we let $I \to (i, a)$ and set the subtraction constant a equal to minus one, giving

$$2p^+ p^- = \frac{1}{\alpha'}\Big(\alpha' p^i p^i + \frac{1}{2}\alpha_0^a \alpha_0^a + \sum_{n=1}^{\infty}\Big[\alpha_{-n}^i \alpha_n^i + \alpha_{-n}^a \alpha_n^a\Big] - 1\Big). \tag{15.49}$$

We therefore have

$$M^2 = 2p^+p^- - p^i p^i = \frac{1}{2\alpha'}\alpha_0^a\alpha_0^a + \frac{1}{\alpha'}\Big(\sum_{n=1}^{\infty}\Big[\alpha_{-n}^i\alpha_n^i + \alpha_{-n}^a\alpha_n^a\Big] - 1\Big). \tag{15.50}$$

Using the explicit value of α_0^a in (15.47), we finally obtain

$$M^2 = \Big(\frac{\bar{x}_2^a - \bar{x}_1^a}{2\pi\alpha'}\Big)^2 + \frac{1}{\alpha'}(N^\perp - 1), \tag{15.51}$$

where

$$N^\perp = \sum_{n=1}^{\infty}\sum_{i=2}^{p} n a_n^{i\,\dagger} a_n^i + \sum_{m=1}^{\infty}\sum_{a=p+1}^{d} m a_m^{a\dagger} a_m^a. \tag{15.52}$$

The first term on the right-hand side of (15.51) is a new contribution to the mass-squared of the states. Since the string tension is $T_0 = 1/(2\pi\alpha')$, this term is simply the square of the energy of a classical static string stretched between the two D-branes. It is reasonable to find that the mass-squared operator is altered by the addition of this constant. The constant vanishes precisely when the branes coincide.

Let us now consider the ground states. In fact, let us examine the ground states from each of the four open string sectors available in this D-brane configuration. The momentum labels of these states are the same for each sector: p^+ and $\vec{p}$. To distinguish the various sectors, we include as additional ground-state labels two integers $[ij]$, each of which can take the value one or two. The first integer denotes the brane on which the $\sigma = 0$ endpoint lies, and the second integer denotes the brane on which the $\sigma = \pi$ endpoint lies. In short, the open strings in the $[ij]$ sector extend from brane i to brane j. The ground states are written as $|p^+, \vec{p}\,;\,[ij]\rangle$, and they are of four types:

$$|p^+, \vec{p}\,;\,[11]\rangle\,, \quad |p^+, \vec{p}\,;\,[22]\rangle\,, \quad |p^+\vec{p}\,;\,[12]\rangle\,, \quad |p^+, \vec{p}\,;\,[21]\rangle. \tag{15.53}$$

The states of open strings in the $[ij]$ sector are constructed from oscillators acting on $|p^+, \vec{p}\,;\,[ij]\rangle$. The states take the form indicated in (15.31), with the exception that the ground state is replaced by $|p^+, \vec{p}\,;\,[ij]\rangle$. The oscillators in the four sectors are the same in number and in type, but they are fundamentally different operators. We could label them

with the $[ij]$ labels for clarity, but this is seldom necessary because the ground states carry the sector labels.

Where do the fields corresponding to the [12] string states live? This question is difficult to answer. They are clearly $(p+1)$-dimensional fields, since the momentum structure of the states is the same as the one we had for the states of strings fully attached to a single brane. As far as the stretched strings are concerned, the two D-branes are on a similar footing, so we cannot say that the fields live on any single one of them. In some sense, the fields must live on *both* D-branes. Operationally, the fields are declared to live on some fixed $(p+1)$-dimensional space (not necessarily identified with any of the two D-branes), and are seen to have nonlocal interactions that reflect the fact that the D-branes are separated. The spacetime interpretation of fields that arise from stretched strings appears to require a new way of thinking, the basis of which may be provided by a branch of mathematics called noncommutative geometry.

We continue our discussion of the state space by giving a list and a detailed description of the fields which comprise the two lowest levels of the stretched-string state space. Just as we did for the single brane, we will determine whether the states are scalars or vectors with respect to the $(p+1)$-dimensional Lorentz symmetry. The simplest states are the ground states:

$$|p^+, \vec{p}\,; [12]\rangle\,, \quad M^2 = -\frac{1}{\alpha'} + \Big(\frac{\bar{x}_2^a - \bar{x}_1^a}{2\pi\alpha'}\Big)^2. \tag{15.54}$$

If the separation between the branes vanishes, these states are tachyon states of the usual mass-squared. If the branes are separated, the mass-squared gets a positive contribution. In fact, for the critical separation

$$|\bar{x}_2^a - \bar{x}_1^a| = 2\pi\sqrt{\alpha'}, \tag{15.55}$$

the ground states represent a massless scalar field. For larger separations, the ground states represent a massive scalar field.

The next states have one oscillator acting on them. Assume, until stated otherwise, that the separation between the branes is nonzero. If the oscillator acting on the ground states arises from a coordinate normal to the brane we have

$$a_1^{a\dagger}\,|p^+, \vec{p}\,; [12]\rangle\,, \quad a = p+1, \ldots, d, \quad M^2 = \Big(\frac{\bar{x}_2^a - \bar{x}_1^a}{2\pi\alpha'}\Big)^2. \tag{15.56}$$

For any momenta, these are $(d-p)$ massive states. Since the index a is not a Lorentz index for the $(p+1)$-dimensional spacetime, these states are Lorentz scalars. We therefore get $(d-p)$ massive scalar fields. If the oscillator arises from a coordinate tangent to the brane we have

$$a_1^{i\,\dagger}|p^+, \vec{p}\,; [12]\rangle\,, \quad i = 2, \ldots, p, \quad M^2 = \Big(\frac{\bar{x}_2^a - \bar{x}_1^a}{2\pi\alpha'}\Big)^2. \tag{15.57}$$

For any momenta, these are $(p+1) - 2 = p - 1$ massive states. Moreover, they carry an index corresponding to the $(p+1)$-dimensional spacetime. We might think that these states make up a massive Maxwell gauge field, but this is not exactly right.

A massive gauge field has more degrees of freedom than a massless gauge field. The results of Problem 10.7 indicate that a massive gauge field has *one more state* than a massless gauge field for each allowed value of the momentum. In a D-dimensional spacetime, a massless gauge field has $D-2$ states for each p_μ which satisfies $p^2=0$, while a massive gauge field has $D-1$ states for each p_μ which satisfies $p^2+m^2=0$. Therefore, in the case at hand, one of the scalar states in (15.56) must join the $(p-1)$ states in (15.57) to form the massive vector. At the end we have one massive vector and $(d-p-1)$ massive scalars.

Can we guess which scalar state in (15.56) becomes part of the massive gauge field? If $p=d-1$ the answer is simple. In that case, the D-branes are separated along one coordinate, and there is only one scalar in (15.56). The scalar uses the oscillator labeled with the direction along which the branes are separated. When $p<d-1$, there is more than one state in (15.56). The scalar state that becomes part of the vector is the linear combination

$$\sum_a (\bar{x}_2^a - \bar{x}_1^a)\, a_1^{a\dagger}\, |p^+, \vec{p}\,; [12]\rangle. \tag{15.58}$$

Of all directions normal to the D-branes, the direction defined by the spatial vector with components $\bar{x}_2^a - \bar{x}_1^a$ is unique: it takes us from one brane to the other one. To visualize this, think of two parallel D1-branes in three-dimensional space. There are clearly many normal directions that do not take us from one brane to the other, and just one direction that does. The educated guess (15.58) can be proven to be the correct one.

We obtain a very interesting situation in the limit as the separation between the branes goes to zero. Even though the D-branes are then coincident, they are still distinguishable and we still have the four open string sectors. The massless open string states which represent strings extending from brane one to brane two include a massless gauge field and $(d-p)$ massless scalars. This is the same field content as that of a sector where strings begin and end on the same D-brane. When the two D-branes coincide we therefore get a total of four massless gauge fields. These gauge fields actually interact with one another – in the string picture they do so by the process of joining endpoints. Theories of interacting gauge fields are called Yang–Mills theories. They were discovered in the 1950s and later on used successfully to build the theories of electroweak and strong interactions. On the world-volume of two coincident D-branes we indeed get a $U(2)$ Yang–Mills theory. More precisely, we get a $U(2)$ Yang–Mills theory with some additional interactions that become negligible at low energies. The two in $U(2)$ is there precisely because we have two coincident D-branes. The meaning of $U(2)$ will be discussed below.

Suppose that we have N Dp-branes. This time the sectors will be labeled by pairs $[ij]$, where i and j are integers that run from 1 to N. The $[ij]$ sector consists of open strings that start on the ith brane and end on the jth brane. It is clear that there are N^2 sectors. In this setup, string interactions can be visualized neatly. In a typical process, a first open string joins with a second open string to form a new open string. To do so, the *end* of the first string ($\sigma=\pi$) joins with the *beginning* of the second string ($\sigma=0$). The new string

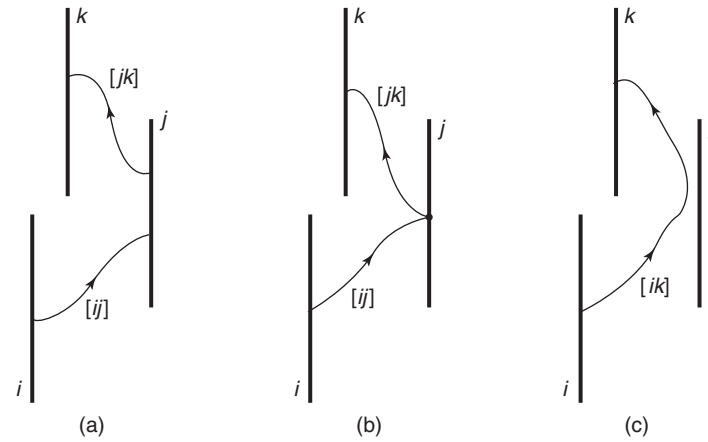

Fig. 15.2 **(a) Three D-branes, labeled *i*, *j*, and *k*, and strings in the [*ij*] and [*jk*] sectors. (b) The end of the string in the [*ij*] sector meets the beginning of the string in the [*jk*] sector and the interaction takes place. (c) The resulting string in the [*ik*] sector.**

begins at the beginning of the first string and ends at the end of the second string. If the open strings are stretched between D-branes, a first string from the $[ij]$ sector can be joined by a second open string from the $[jk]$ sector to give a "product" open string in the $[ik]$ sector. This interaction is possible since both the end of the first string and the beginning of the second string lie on the same D-brane. The physical process can be imagined to take place as in Figure 15.2. The result is a single string, which does not remain attached to the j D-brane since the joining point is no longer an endpoint. The new string belongs to the $[ik]$ sector. We write this possible interaction as

$$[ij] * [jk] = [ik], \quad j \text{ not summed.} \tag{15.59}$$

If the N Dp-branes are coincident, the N^2 sectors result in N^2 interacting massless gauge fields. This defines a $U(N)$ Yang–Mills theory on the world-volume of the N coincident D-branes:

$$N \text{ coincident D-branes carry } U(N) \text{ massless gauge fields.} \tag{15.60}$$

In fact, the full spectrum of the open string theory consists of N^2 copies of the spectrum of a single Dp-brane.

If we have a single brane, (15.60) tells us that we get a $U(1)$ Yang–Mills theory. In fact, $U(1)$ Yang–Mills theory is Maxwell theory, so (15.60), for $N = 1$, is consistent with (15.35). Here $U(1)$ denotes a group: the elements of the group are complex numbers of unit length and group multiplication is just multiplication. The $U(1)$ group is relevant because, at any spacetime point, the gauge parameters in Maxwell theory are actually elements of $U(1)$. So far, our study of Maxwell gauge transformations has made no use of the $U(1)$

group structure, but this group structure is needed to understand the gauge symmetry in the presence of compact spatial dimensions (Chapter 18). For $U(N)$ Yang–Mills theory, $U(N)$ is also a group of symmetries: the elements of the group are $N \times N$ unitary matrices and group multiplication is matrix multiplication. At any spacetime point, the gauge parameters of $U(N)$ Yang–Mills theory are elements of the group $U(N)$.

Quick calculation 15.1 Recall that a group is a set which is closed under an associative multiplication; it contains an identity element, and each element has a multiplicative inverse. Verify that $U(1)$ and $U(N)$, as described above, are groups.

The discrete labels i, j used to label the branes and the various open string sectors are sometimes called *Chan–Paton* indices. These indices were introduced during the early stages of string theory, long before D-branes were known, as an algebraic device to obtain Yang–Mills theories from open strings. With the discovery of D-branes it became clear that the Chan–Paton indices are simply labels of D-branes in a multi-D-brane configuration.

The appearance of Yang–Mills theories on the world-volume of a D-brane configuration is of great relevance because Yang–Mills theories are used to describe the Standard Model of particle physics. The electroweak theory is described by a $U(2)$ Yang–Mills theory. The four gauge bosons of this theory include the photon γ, the W^+, the W^-, and the Z^0. The latter three are massive gauge bosons. The mechanism by which massless gauge fields become massive is known as the Higgs mechanism of field theory. A possible D-brane realization of the Higgs mechanism is obtained by separating D-branes which, when coincident, give the corresponding massless gauge particles. If we have two coincident D3-branes we obtain a $U(2)$ Yang–Mills theory, with four massless gauge fields living on the four-dimensional world-volume of the branes. Is this a good model for the electroweak gauge theory? Not quite. If we separate the D3-branes to give mass to some of the gauge bosons, two of them acquire a mass – the two arising from the stretched strings – and two remain massless. In the electroweak gauge theory only one gauge field remains massless. A more sophisticated D-brane configuration is needed to produce a model of the electroweak theory. A setup with intersecting D-branes that can be used to construct a particle physics model will be examined in Section 21.1. The construction of semi-realistic models is discussed in Section 21.4.

15.4 Strings between parallel Dp- and Dq-branes

In this section we examine the configuration of two parallel D-branes with different dimensionality. Let p and q be two integers which satisfy $1 \leq q < p \leq 25$, and consider a configuration consisting of a Dp-brane and a Dq-brane. We assume $p > q$, since the case $p = q$ was already considered. The branes are coincident if the Dq-brane world-volume is a subset of the Dp-brane world-volume. We take the branes to be parallel. This means the same as what we mean when we say that a line is parallel to a plane: there is a plane

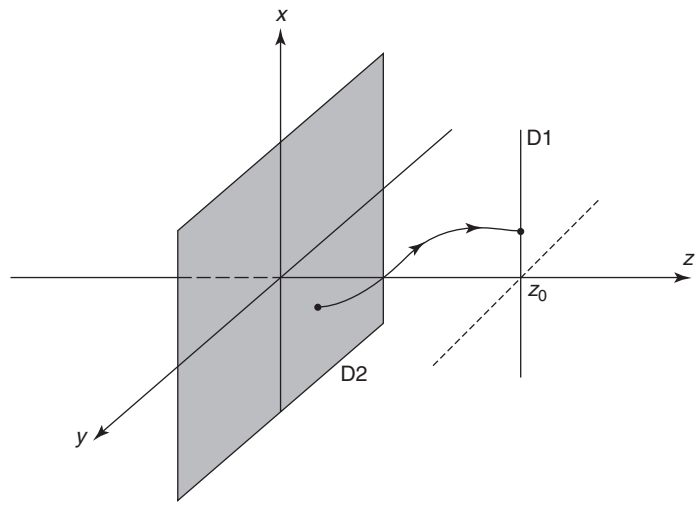

Fig. 15.3 **A D2-brane stretched on the (x, y) plane and a parallel D1-brane stretched along the x axis and located at $y = 0$, $z = z_0$. Also shown is an open string going from the D2-brane to the D1-brane. For such a string, the string coordinate Y is of ND type.**

parallel to the given plane which contains the line. Thus, if the Dp-brane and Dq-brane are separate, there is a p-dimensional hyperplane parallel to the Dp-brane that contains the Dq-brane.

A case that can be easily visualized is that of a D2-brane and a D1-brane, as illustrated in Figure 15.3. The D2-brane stretches along the x and y directions and is located at $z = 0$. The D1-brane stretches along the x direction and is located at $y = 0$ and at $z = z_0$. This D1-brane is parallel to the D2-brane, and they are coincident only when $z_0 = 0$. Table 15.1 summarizes the relevant spatial information about the D-branes. A dash (—) indicates a direction along the D-brane, and a bullet (•) indicates a direction normal to the D-brane. The coordinate x is a common tangential direction. The coordinate y is a mixed direction: one brane extends along this direction while the other does not. The z coordinate is a common normal direction. More generally, for the Dp-, Dq-brane configuration we have

$$\underbrace{x^0,\, x^1,\, \ldots, x^q}_{\text{common tangential coordinates}} \quad \underbrace{x^{q+1},\, x^{q+2},\, \ldots,\, x^p}_{\text{mixed coordinates}} \quad \underbrace{x^{p+1},\, x^{p+2},\, \ldots,\, x^d}_{\text{common normal coordinates}} . \tag{15.61}$$

There are $(q + 1)$ common tangential coordinates (all the world-volume coordinates of the Dq-brane, including time), $(p - q)$ directions that are tangential to the Dp-brane and normal to the Dq-brane, and $(d - p)$ common normal directions.

We have already studied strings that begin and end on the same D-brane, so we focus here on the strings that go from one D-brane to the other. For definiteness, consider the strings that stretch from the Dp-brane to the Dq-brane. We have partial knowledge about these strings: the common tangential coordinates are NN, and the common normal coordinates are DD. We have already studied these two types of coordinates. The mixed coordinates, which are tangential to one of the branes and normal to the other, are new. In

Table 15.1 D-brane configuration and boundary conditions

Coordinate	x	y	z
D2	–	–	•
D1	–	•	•
[D2 , D1]	NN	ND	DD

Note. In the second and third rows a dash – indicates coordinates along which the D-brane stretches and a bullet • indicates coordinates normal to the brane. In the last row we indicate the boundary conditions for open strings belonging to the sector [D2, D1] of strings that stretch from the D2-brane to the D1-brane.

our D2-, D1-brane example the y direction is mixed. For a string that stretches from the D2-brane to the D1-brane, the string coordinate Y associated with y is N at $\sigma = 0$, since y is tangential to the D2-brane, and D at $\sigma = \pi$, since y is normal to the D1-brane; in short, Y is an ND coordinate. For a string that stretches from the D1-brane to the D2-brane, Y is a DN coordinate.

In the Dp-, Dq-brane configuration, the mixed spatial coordinates were listed in (15.61). For open strings that stretch from the Dp-brane to the Dq-brane, the analogously labeled string coordinates satisfy a Neumann boundary condition on the starting Dp-brane and a Dirichlet boundary condition on the ending Dq-brane; they are ND coordinates. The full set of string coordinates splits into

$$\underbrace{X^0,\, X^1,\, \ldots,\, X^q}_{\text{NN coordinates}}\ \underbrace{X^{q+1},\, X^{q+2},\, \ldots,\, X^p}_{\text{ND coordinates}}\ \underbrace{X^{p+1},\, X^{p+2},\, \ldots,\, X^d}_{\text{DD coordinates}}. \tag{15.62}$$

In the light-cone, we use three types of indices to label the string coordinates:

$$\underbrace{X^+, X^-, \{X^i\}}_{\text{NN}}\quad \underbrace{\{X^r\}}_{\text{ND}}\quad \underbrace{\{X^a\}}_{\text{DD}}, \tag{15.63}$$

where

$$i = 2, \ldots, q, \quad r = q+1, \ldots, p, \quad \text{and} \quad a = p+1, \ldots, d. \tag{15.64}$$

We think of the Dp-brane as the first brane and the Dq-brane as the second brane. The position of the Dp-brane is specified by the coordinates $\bar{x}_1^a$, and the position of the Dq-brane is specified by the coordinates $\bar{x}_2^r$ and $\bar{x}_2^a$. In our D2-, D1-brane example of Figure 15.3, the role of $\bar{x}_2^r$ is played by the y coordinate of the D1-brane. This coordinate can be set to zero by a suitable choice of axes.

Let us begin our analysis of the ND coordinates X^r. The boundary conditions are

$$\frac{\partial X^r}{\partial \sigma}(\tau,\sigma)\Big|_{\sigma=0} = 0\,, \quad X^r(\tau,\sigma)\Big|_{\sigma=\pi} = \bar{x}_2^r. \tag{15.65}$$

The $\bar{x}_2^r$ coordinates can be set equal to zero by a suitable choice of axes, but we will not do so. As opposed to the coordinate differences $\bar{x}_1^a - \bar{x}_2^a$ that define the separation between the D-branes, the $\bar{x}_2^r$ will not play any significant role. Consider now the usual expansion

$$X^r(\tau,\sigma) = \frac{1}{2}\Big(f^r(\tau+\sigma) + g^r(\tau-\sigma)\Big). \tag{15.66}$$

The boundary condition at $\sigma = 0$ gives us

$$f^{r\prime}(u) = g^{r\prime}(u) \;\rightarrow\; g^r(u) = f^r(u) + c_0^r. \tag{15.67}$$

Bearing in mind that the second boundary condition will set X^r equal to $\bar{x}_2^r$ at $\sigma = \pi$, we choose $c_0^r = 2\bar{x}_2^r$ so that

$$X^r(\tau,\sigma) = \bar{x}_2^r + \frac{1}{2}\Big(f^r(\tau+\sigma) + f^r(\tau-\sigma)\Big). \tag{15.68}$$

The condition at $\sigma = \pi$ then gives

$$f^r(u+2\pi) = -f^r(u). \tag{15.69}$$

The function $f^r(u)$ goes into *minus* itself when its argument increases by 2π. We encountered a similar situation for the twisted sector of the $\mathbb{R}^1/\mathbb{Z}_2$ orbifold (Section 13.6) and for Neveu–Schwarz fermions (Section 14.4). What are needed are trigonometric functions with half-integer moding:

$$f^r(u) = \sum_{n\in\mathbb{Z}^+_{\text{odd}}} \Big[f_n^r \cos\Big(\frac{nu}{2}\Big) + h_n^r \sin\Big(\frac{nu}{2}\Big)\Big]. \tag{15.70}$$

Substituting back into (15.68) and relabeling the expansion coefficients, we get

$$X^r(\tau,\sigma) = \bar{x}_2^r + \sum_{n\in\mathbb{Z}^+_{\text{odd}}} \Big[A_n^r \cos\Big(\frac{n\tau}{2}\Big) + B_n^r \sin\Big(\frac{n\tau}{2}\Big)\Big] \cos\Big(\frac{n\sigma}{2}\Big). \tag{15.71}$$

This is our expansion of the ND coordinates. To proceed with the quantization we define oscillators with half-integer moding. Another useful guide is the desired simplicity of $\dot{X}^r \pm X^{r\prime}$. We are thus led to write

$$X^r(\tau,\sigma) = \bar{x}_2^r + i\sqrt{2\alpha'} \sum_{n\in\mathbb{Z}_{\text{odd}}} \frac{2}{n}\, \alpha^r_{\frac{n}{2}}\, e^{-i\frac{n}{2}\tau} \cos\Big(\frac{n\sigma}{2}\Big), \tag{15.72}$$

where the sum runs over both positive and negative odd integers. The factor of i in front of the sum is necessary so that the Hermiticity of X^r imposes the standard Hermiticity property on the oscillators:

$$\big(\alpha^r_{\frac{n}{2}}\big)^\dagger = \alpha^r_{-\frac{n}{2}}. \tag{15.73}$$

The $\bar{x}_2^r$ are constants and do not become operators. There are no zero modes in the expansion of X^r, and therefore ND coordinates carry no average momentum. We also record the derivatives

$$\dot{X}^r \pm X^{r\prime} = \sqrt{2\alpha'} \sum_{n\in\mathbb{Z}_{\text{odd}}} \alpha^r_{\frac{n}{2}}\, e^{-i\frac{n}{2}(\tau\pm\sigma)}, \tag{15.74}$$

which are indeed of the expected form. This expansion reminds us of similar ones (13.106) for the twisted sector orbifold coordinate X.

The (nontrivial) commutation relations for the string coordinates take the form

$$[X^r(\tau,\sigma),\dot{X}^s(\tau,\sigma')] = i(2\pi\alpha')\delta(\sigma-\sigma')\delta^{rs}. \tag{15.75}$$

This equation implies that we can use the appropriate version of (12.30):

$$\left[\,(\dot{X}^r \pm X^{r\prime})(\tau,\sigma)\,,\;(\dot{X}^s \pm X^{s\prime})(\tau,\sigma')\,\right] = \pm 4\pi\alpha'\, i\eta^{rs}\frac{d}{d\sigma}\delta(\sigma-\sigma'). \tag{15.76}$$

Since the expansions (15.74) take the standard form, equation (12.40) also holds, with minor modifications, for $\sigma,\sigma' \in [0,2\pi]$:

$$\sum_{m',n'\in\mathbb{Z}_{\rm odd}} e^{-i\frac{m'}{2}(\tau+\sigma)}e^{-i\frac{n'}{2}(\tau+\sigma')}\left[\alpha^r_{\frac{m'}{2}}\,,\alpha^s_{\frac{n'}{2}}\right] = 2\pi\, i\eta^{rs}\frac{d}{d\sigma}\delta(\sigma-\sigma'). \tag{15.77}$$

The commutators are extracted just like we did below (13.107) and the answer is

$$\left[\,\alpha^r_{\frac{m}{2}}\,,\alpha^s_{\frac{n}{2}}\,\right] = \tfrac{m}{2}\,\delta^{rs}\,\delta_{m+n,0}. \tag{15.78}$$

This is the expected form of the commutation relations.

Let us now calculate the mass-squared operator. This operator receives contributions from all the coordinates in this sector: the NN, the ND, and the DD coordinates. This is clear from (15.8) since the original light-cone index I now runs over i, r, and a labels. Given that the linear combination of derivatives (15.74) takes the standard form, the contribution from the ND coordinates takes a familiar form. The formula for $2p^+p^-$ can be obtained by a minor modification of equation (15.49):

$$2p^+\,p^- = \frac{1}{\alpha'}\Big(\alpha' p^i p^i + \frac{1}{2}\,\alpha^a_0\,\alpha^a_0 + \sum_{n=1}^{\infty}\Big[\alpha^i_{-n}\,\alpha^i_n + \alpha^a_{-n}\,\alpha^a_n\Big] + \sum_{m\in\mathbb{Z}^+_{\rm odd}} \alpha^r_{-\frac{m}{2}}\alpha^r_{\frac{m}{2}} + a\Big)\,. \tag{15.79}$$

In writing this equation we have restored the ordering constant a, which, as you recall, arises heuristically by ordering the oscillators in $L_0^\perp$ (see (12.102), (12.107), and (12.110)). Since all the oscillators for both the NN and DD directions are integrally moded, their normal-ordering constants are the same and equal to $\frac{1}{2}(\frac{-1}{12}) = -\frac{1}{24}$. With a total of 24 transverse light-cone coordinates, if we only have NN and DD coordinates we get $a=-1$. The ND coordinates, however, give a different contribution. The sum that must be rearranged in this case is

$$\frac{1}{2}\sum_{m\in\mathbb{Z}_{\rm odd}} \alpha^r_{-\frac{m}{2}}\alpha^r_{\frac{m}{2}} = \sum_{m\in\mathbb{Z}^+_{\rm odd}} \alpha^r_{-\frac{m}{2}}\alpha^r_{\frac{m}{2}} + \frac{1}{2}\sum_{m\in\mathbb{Z}^+_{\rm odd}}\left[\alpha^r_{\frac{m}{2}}\,,\alpha^r_{-\frac{m}{2}}\right]. \tag{15.80}$$

The first term on the right-hand side is the one that appears in (15.79), and the second term is the ordering contribution. Since we have $(p-q)$ ND coordinates, the ordering constant is

$$\frac{1}{2}\sum_{m\in\mathbb{Z}^+_{\rm odd}}\left[\alpha^r_{\frac{m}{2}}\,,\alpha^r_{-\frac{m}{2}}\right] = \frac{1}{4}\,(p-q)\sum_{m\in\mathbb{Z}^+_{\rm odd}} m = \frac{1}{48}\,(p-q), \tag{15.81}$$

where we used (15.78) and recalled that the sum over positive odd integers is equal to $\frac{1}{12}$. This calculation shows that each ND coordinate contributes $+\frac{1}{48}$ to the ordering constant in $\alpha' M^2$. A DN string coordinate will contribute the same amount. In summary, the normal ordering contributions to $\alpha' M^2$ for the various string coordinates are

$$a_{\rm NN} = a_{\rm DD} = -\frac{1}{24}\,, \quad a_{\rm ND} = a_{\rm DN} = \frac{1}{48}. \tag{15.82}$$

Returning to the problem at hand, the total ordering constant a is given by (15.81) plus the contribution of the $(24-(p-q))$ coordinates that are either NN or DD:

$$a = -\frac{1}{24}\big(24-(p-q)\big) + \frac{1}{48}(p-q) = -1 + \frac{1}{16}(p-q). \tag{15.83}$$

With this information we can now find M^2. Following the same steps as in (15.50),

$$M^2 = \Big(\frac{\bar{x}_2^a - \bar{x}_1^a}{2\pi\alpha'}\Big)^2 + \frac{1}{\alpha'}\Big(N^\perp - 1 + \frac{1}{16}(p-q)\Big), \tag{15.84}$$

where

$$N^\perp = \sum_{n=1}^{\infty}\sum_{i=2}^{q} n\, a_n^{i\,\dagger} a_n^i + \sum_{k\in\mathbb{Z}^+_{\rm odd}}\sum_{r=q+1}^{p} \frac{k}{2}\, a_{\frac{k}{2}}^{r\dagger} a_{\frac{k}{2}}^{r} + \sum_{m=1}^{\infty}\sum_{a=p+1}^{d} m\, a_m^{a\dagger} a_m^a. \tag{15.85}$$

This formula for M^2 incorporates all the effects we have discussed: a shifted ordering constant, a contribution from stretched strings, and a number operator that includes contributions from NN, DD, and ND coordinates.

Let us examine the state space and the fields associated with the two lowest mass levels. The ground states are labeled as

$$|p^+, \vec{p}\,; [12]\rangle\,, \quad \vec{p} = (p^2, \ldots, p^q). \tag{15.86}$$

The momentum labels on the states indicate that the corresponding fields live in a $(q+1)$-dimensional spacetime. Roughly, they live on the world-volume of the Dq-brane, the brane of lower dimensionality. The general rule is clear: the spacetime dimensionality of the fields which arise in any given sector equals the number of NN string coordinates in the sector. The state space is built by letting the three types of oscillators – $a_p^{i\dagger}$, $a_{k/2}^{r\dagger}$, and $a_m^{a\dagger}$ – act on the ground states.

The ground states have $N^\perp = 0$ and correspond to a single scalar field on the Dq-brane. This scalar is in general massive, but it can be tachyonic or massless depending on the separation of the branes and the value of $p-q$. Assume, for simplicity, that the branes coincide. If, additionally, $p-q=16$, then the scalar is massless. The next states are of the form

$$a_{\frac{1}{2}}^{r\dagger}|p^+, \vec{p}\,; [12]\rangle\,, \quad N^\perp = 1/2. \tag{15.87}$$

These states give rise to $(p-q)$ scalar fields, since the index r does not correspond to a world-volume direction on the Dq-brane. All other states are necessarily massive since

they have $N^\perp \geq 1$, and this together with $p > q$ implies that $M^2 > 0$. In particular, we do not find any massless gauge fields.

Problems

Problem 15.1 A Dp-brane with orientifolds.

In this problem (a sequel to Problem 13.6), we study the effects of orientifolds on open strings.

The space-filling O25-plane truncates the spectrum down to the set of states that are invariant under the operation Ω which reverses the orientation of strings. When we have a Dp-brane, Ω acts on the open string coordinates as follows:

$$\Omega\, X^a(\tau,\sigma)\, \Omega^{-1} = X^a(\tau, \pi - \sigma), \tag{1}$$

$$\Omega\, X^i(\tau,\sigma)\, \Omega^{-1} = X^i(\tau, \pi - \sigma). \tag{2}$$

As usual, we demand that $\Omega\, x_0^-\, \Omega^{-1} = x_0^-$ and $\Omega\, p^+\, \Omega^{-1} = p^+$.

(a) Give the Ω action on the oscillators α_n^a and α_n^i. What is the expected Ω action on α_n^-? Does it work out?

(b) Assume that the ground states $|p^+, \vec{p}\,\rangle$ are Ω invariant. Find the states of the theory for $N^\perp \leq 2$. As you will see, some massless states survive. Interpret these states along the lines of the discussion below (15.37).

Replace the O25-plane by an Op-plane coincident with the Dp-brane at $\bar{x}^a = 0$. Let Ω_p denote the operator for which this theory keeps only the states with $\Omega_p = +1$.

(c) How should equations (1) and (2) change when Ω is replaced by Ω_p? Give the Ω_p action on the oscillators α_n^a and α_n^i.

(d) Describe the full spectrum of the theory as a simple truncation of the Dp-brane spectrum. You will find no massless scalars in this case. What does this suggest regarding the possible motions of the Dp-brane?

Problem 15.2 String products and orientation reversing symmetries.

Equation (15.59) tells how open string *sectors* combine under interactions. The same product notation can be used for strings. By

$$|A\rangle * |B\rangle \tag{1}$$

we mean the string state that is obtained when a string in state $|A\rangle$ interacts with a string in state $|B\rangle$. The string product must obey the rule of sectors: the state in (1) must belong to the sector $[A] * [B]$, where $[A]$ and $[B]$ denote the sectors where string states $|A\rangle$ and $|B\rangle$ belong, respectively.

Use pictures of strings A and B to motivate the equations

$$\Omega\,(|A\rangle * |B\rangle) = (\Omega|B\rangle) * (\Omega|A\rangle), \tag{2}$$

$$\Omega_p\,(|A\rangle * |B\rangle) = (\Omega_p|B\rangle) * (\Omega_p|A\rangle). \tag{3}$$

Here Ω is string orientation reversal and Ω_p is orientifolding (orientation reversal plus reflection about a set of coordinates).

Problem 15.3 N coincident Dp-branes and orientifolds.

Let N coincident Dp-branes coincide with an Op-plane, all of them located at $\bar{x}^a = 0$. The orientifolding symmetry Ω_p, as usual, includes reflection of the coordinates normal to the orientifold and simultaneous orientation reversal of strings. Assume that the reflection of coordinates leaves each of the Dp-branes invariant (as opposed to mapping them into each other). The states of the theory are those for which $\Omega_p = +1$.

(a) Explain why it is reasonable to postulate that

$$\Omega_p|p^+, \vec{p}\,; [ij]\rangle = |p^+, \vec{p}\,; [ji]\rangle.$$

What are the ground states of the theory? How many are there?

(b) Describe the full open string spectrum of the theory in terms of the spectrum of a single Dp-brane. Check that for $N = 1$ you reproduce the result of Problem 15.1 (d).

Problem 15.4 Separated Dp-branes and an Op-plane.

We have learned that an orientifold acts as a kind of mirror. If we are to have D-branes that do not coincide with an orientifold, then there must be mirror D-branes at the reflected points. Therefore, to analyze the theory of Dp-branes off an orientifold Op-plane we begin with N Dp-branes and N mirror Dp-branes at the reflected positions. We must then define the orientifold action on all the states of the theory of $2N$ Dp-branes. Finally, we use this action to truncate down to the invariant states, obtaining in this way the states of the orientifold theory.

Consider the situation illustrated in Figure 15.4, where we show the configuration as seen in a plane spanned by two coordinates normal to the branes and the orientifold. The N Dp branes are labeled $1, 2, \ldots, N$, and the mirror images are labeled $\bar{1}, \bar{2}, \ldots, \bar{N}$. Two strings are exhibited: one in the [24] sector and the other in the $[1\bar{1}]$ sector.

(a) Show the two strings obtained by the orientifold symmetry. Since the arguments $p^+, \vec{p}$ of the ground states are always present, let us omit them for brevity. The ground states are of four types:

$$|[ij]\rangle\,, \quad |[i\,\bar{j}]\rangle\,, \quad |[\bar{i}\,j]\rangle\,, \quad |[\bar{i}\,\bar{j}]\rangle. \tag{1}$$

Each class contains N^2 ground states since i and j run from 1 to N and $\bar{i}$ and $\bar{j}$ run from $\bar{1}$ to $\bar{N}$. Define an expected action of Ω_p on the ground states in (1). Show that your choice satisfies $\Omega_p^2 = 1$ acting on the ground states.

(b) What are the possible interactions between strings in the four types of sectors built on the states (1)? Write your answers using the notation of (15.59).

(c) It is a fact about string interactions that the string product of ground states gives states that have a component along a ground state. Thus, for example,

$$|[i\,\bar{j}]\rangle * |[\bar{j}\,k]\rangle = |[i\,k]\rangle + \cdots. \tag{2}$$

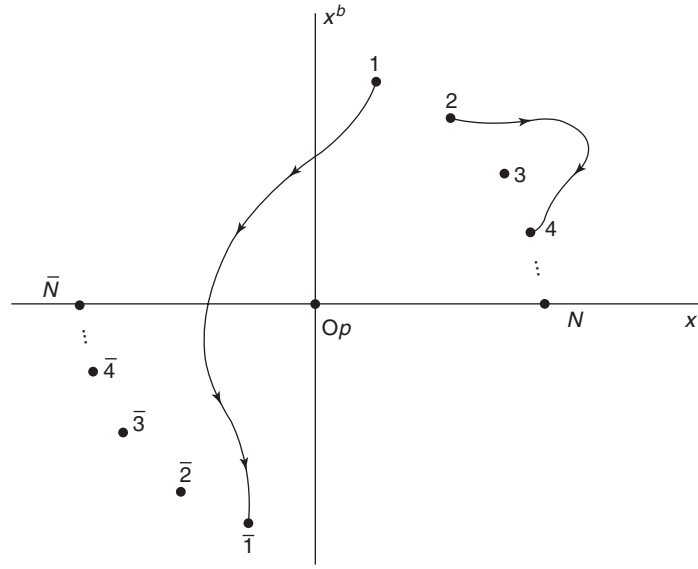

Fig. 15.4 **Problem 15.4. A set of N Dp-branes together with a set of N image Dp-branes. The orientifold Op plane is at the origin.**

Write the other possible ground state products. Test the consistency of your definition of Ω_p by acting with Ω_p on both sides of the equations giving ground state products. To act on products, use equation (3) of Problem 15.2.

(d) Find the Ω_p action on the α_n^i oscillators using $\Omega_p X^i(\tau,\sigma)\Omega_p^{-1} = X^i(\tau,\pi-\sigma)$. Since the strings are stretched along, or have nonzero values for, the x^a coordinates, an equation of the type $\Omega_p X^a(\tau,\sigma)\Omega_p^{-1} = -X^a(\tau,\pi-\sigma)$ cannot be fully implemented. Using (15.20) for X^a, for example, we would need $\Omega_p \bar{x}^a \Omega_p^{-1} = -\bar{x}^a$, which cannot hold since the $\bar{x}^a$ are numbers. A legal derivation of the Ω_p action on the α_n^a oscillators can be obtained by requiring $\Omega_p \dot{X}^a(\tau,\sigma)\Omega_p^{-1} = -\dot{X}^a(\tau,\pi-\sigma)$. Verify that for any arbitrary product R of oscillators of both types

$$\Omega_p \, R \, \Omega_p^{-1} = (-1)^{N^\perp} R, \tag{3}$$

where $N^\perp$ is the total number of R.

(e) Describe the orientifold spectrum in terms of the spectrum of a single Dp-brane. For this, consider an arbitrary product R of oscillators and build the general states

$$\sum_{ij}\Big(r_{ij}R|[ij]\rangle + r_{i\bar{j}}R|[i\bar{j}]\rangle + r_{\bar{i}j}R|[\bar{i}j]\rangle + r_{\bar{i}\bar{j}}R|[\bar{i}\bar{j}]\rangle\Big), \tag{4}$$

where $r_{ij}, r_{i\bar{j}}, r_{\bar{i}j}$, and $r_{\bar{i}\bar{j}}$ are four N by N matrices. Find the conditions that Ω_p invariance imposes on these matrices. There are two cases to consider, depending on the number $N^\perp$ of R. You should find that for $N^\perp$ odd there are $N(2N-1)$ linearly independent states in (4). For $N^\perp$ even there are $N(2N+1)$ linearly independent states.

Since the gauge fields arise from $N^\perp = 1$ states, there are $N(2N-1)$ of them. This is the number of entries in a $2N$-by-$2N$ antisymmetric matrix. The interacting theory of such gauge fields is an $SO(2N)$ Yang–Mills gauge theory (SO stands for special orthogonal). The $SO(10)$ gauge theory, for example, can be used to build a grand unified theory of the strong and electroweak interactions.

Problem 15.5 Separated Dp-branes and a different Op-plane.

The brane setup here is that of Problem 15.4. In part (a) of that problem you defined a simple action of Ω_p on the ground states $|[ij]\rangle$, $|[i\ \bar{j}]\rangle$, $|[\bar{i}\ j]\rangle$, and $|[\bar{i}\ \bar{j}]\rangle$. Find an alternative Ω_p action where some of the relations have a minus sign: $\Omega_p|[\dots]\rangle = \pm|[\dots]\rangle$. Test its consistency by verifying that $\Omega_p^2 = 1$ on ground states and that the products of ground states are compatible with the Ω_p action, as you tested in part (c) of Problem 15.4. Determine the spectrum of this variant orientifold theory.

The gauge fields arise from $N^\perp = 1$, and you will find that there are $N(2N+1)$ of them. The interacting theory of such gauge fields is a $USp(2N)$ Yang–Mills gauge theory (USp stands for unitary symplectic).

Problem 15.6 DN string coordinates.

In Section 15.4 we considered strings that stretch from a Dp- to a Dq-brane, focusing on the coordinates X^r which satisfy ND boundary conditions. Consider now strings that stretch from the Dq- to the Dp-brane. For such strings the coordinates X^r are of DN type.

(a) Write the boundary conditions satisfied by the X^r coordinates and use them to derive a mode expansion along the lines of our result (15.72) for an ND coordinate.

(b) Find also the equations that replace (15.74). Explain briefly why the mass-squared formula (15.84) needs no modification.

(c) If we let $\sigma \to \pi - \sigma$ in (15.72) we automatically get a function with DN boundary conditions. Compare with the mode expansion you found in (a), and explain why the Hermiticity properties are consistent.

Problem 15.7 D1-branes at an angle.

Consider two infinitely long D1-branes stretched on the (x^2, x^3) plane. The *first* brane is defined by $x^3 = 0$, and the *second* brane is at an angle γ measured counterclockwise from the x^2 axis. Let the open string coordinates be $X^2(\tau, \sigma)$ and $X^3(\tau, \sigma)$, and consider only open strings which begin on the first brane and end on the second brane. Determine the boundary conditions satisfied by X^2 and X^3 at $\sigma = 0$ and $\sigma = \pi$.

Problem 15.8 Strings in a configuration with a Dp-brane and a D25-brane.

Consider the full state space of open string theory in a configuration with a Dp-brane and a D25-brane. Assume $1 \le p \le 24$. For each sector of the theory give the M^2 operator, and examine explicitly the states which arise in the two lowest levels, indicating the types of fields they correspond to and where these fields live.

Problem 15.9 A pair of intersecting D22-branes.

We study here a configuration of two D22-branes. One of them, henceforth called brane 1, is defined by $x^{25} = x^{23} = x^{22} = 0$. The other one, brane 2, is defined by $x^{24} = x^{23} = x^{22} = 0$.

Draw a picture of the brane configuration as it appears in the (x^{24}, x^{25}) plane. Construct a table, such as Table 15.1, adding more columns to include all coordinates and more rows to include all sectors. For each sector of the theory give the M^2 operator and examine explicitly the states which arise in the two lowest levels, indicating the types of fields they correspond to and where these fields live.

A basic point: if x^a is a common Dirichlet direction in a configuration of two intersecting D-branes, then the x^a coordinates of the two D-branes must be the same. Explain why.

16 String charge and electric charge

If a point particle couples to the Maxwell field then that particle carries electric charge. Strings couple to the Kalb–Ramond field; therefore, strings carry a new kind of charge – string charge. For a stretched string, string charge can be visualized as a current flowing along the string. Strings can end on D-branes without violating conservation of string charge because the string endpoints carry electric charge and the resulting electric field lines on a D-brane carry string charge. Certain D-branes in superstring theory carry electric charge for Ramond–Ramond fields. If a charged brane is fully wrapped on a compact space, it appears to a lower-dimensional observer as a point particle carrying the electric charge of a Maxwell field that arises from dimensional reduction.

16.1 Fundamental string charge

As we have seen before, a point particle can carry electric charge because there is an interaction which allows the particle to couple to a Maxwell field. The world-line of the point particle is *one*-dimensional and the Maxwell gauge field A_μ carries *one* index. This matching is important. The particle trajectory has a tangent vector $dx^\mu(\tau)/d\tau$, where τ parameterizes the world-line. Because it has one Lorentz index, the tangent vector can be multiplied by the gauge field A_μ to form a Lorentz scalar. Working with natural units ($\hbar = c = 1$), the interaction for a point particle of charge q is written as a term in the action taking the form

$$q \int A_\mu(x(\tau)) \frac{dx^\mu(\tau)}{d\tau} d\tau. \tag{16.1}$$

It is convenient to require that q be dimensionless in natural units. Since the action is also dimensionless in natural units, the gauge field A_μ must carry units of inverse length, or mass: $[A_\mu] = M$. The field strength has units $[F_{\mu\nu}] = M^2$ because it is obtained by differentiation of the gauge field with respect to the spacetime coordinates.

The complete interacting system of the charged particle and the Maxwell field is defined by the action considered in Problem 5.6:

$$S' = -m \int_{\mathcal{P}} ds + q \int_{\mathcal{P}} A_\mu(x) dx^\mu - \frac{1}{4\kappa_0^2} \int d^D x \; F_{\mu\nu} F^{\mu\nu}. \tag{16.2}$$

The first term on the right-hand side is the particle action, and the last term is the action for the Maxwell field. We have included the dimensionful constant κ_0, with units

$[\kappa_0^2] = M^{4-D}$, in order to make this term dimensionless. This constant is necessary whenever $D \neq 4$.

Can a relativistic string be charged? The above argument makes it clear that Maxwell charge is naturally carried by points. Closed strings do not have special points, but open strings have distinguished endpoints. It is therefore plausible that the endpoints of open strings carry electric Maxwell charge. We will show later that this is indeed the case. At the moment, however, we are looking for something fundamentally different. Since electric Maxwell charge is naturally associated with points, we may wonder whether there is some new kind of charge that is naturally associated with strings. For a new kind of charge, we need a new kind of gauge field. Thus, we may ask: is there a field in string theory that is related to the string in the same way that the Maxwell field is related to a particle? The answer is yes. The field is the Kalb–Ramond antisymmetric two-tensor $B_{\mu\nu}(= -B_{\nu\mu})$. This is a massless field that arises in closed string theory (see Section 13.3 and Problem 10.6).

Let us now mimic the logic that led to (16.1). At any point of the string trajectory we have two linearly independent tangent vectors. Indeed, with world-sheet coordinates τ and σ, the two tangent vectors can be chosen to be $\partial X^\mu/\partial\tau$ and $\partial X^\mu/\partial\sigma$. With these two tangent vectors and the two-index field $B_{\mu\nu}$ we can construct a Lorentz scalar:

$$-\int d\tau d\sigma \; \frac{\partial X^\mu}{\partial\tau}\frac{\partial X^\nu}{\partial\sigma} B_{\mu\nu}\left(X(\tau,\sigma)\right). \tag{16.3}$$

This is how the string couples to the antisymmetric Kalb–Ramond field. It is called an *electric* coupling because it is the natural generalization of the electric coupling of a point particle to a Maxwell field. Thus we say that the string carries *electric Kalb–Ramond charge*. The coupling (16.3) must be dimensionless in natural units, so $B_{\mu\nu}$ carries units of inverse length-squared or mass-squared: $[B_{\mu\nu}] = M^2$. This coupling must also be invariant under (τ, σ) reparameterizations, just like the Nambu–Goto string action is. You will see in Problem 16.1 that the antisymmetry of $B_{\mu\nu}$ is necessary to ensure the reparameterization invariance of (16.3). An important point regarding the extent to which reparameterization invariance holds will be addressed later in this section.

Just as (16.2) represents the complete dynamics of a particle and a Maxwell field, the string coupling (16.3) must be supplemented by the string action $S_{\rm str}$ and a term giving dynamics to the $B_{\mu\nu}$ field:

$$S = S_{\rm str} - \frac{1}{2}\int d\tau d\sigma\, B_{\mu\nu}\left(X(\tau,\sigma)\right)\frac{\partial X^{[\mu}}{\partial\tau}\frac{\partial X^{\nu]}}{\partial\sigma} - \frac{1}{6\kappa^2}\int d^D x\, H_{\mu\nu\rho}H^{\mu\nu\rho}. \tag{16.4}$$

Here we have defined the antisymmetrization

$$a^{[\mu}b^{\nu]} \equiv a^\mu b^\nu - a^\nu b^\mu \tag{16.5}$$

and the field strength $H_{\mu\nu\rho}$ associated with $B_{\mu\nu}$:

$$H_{\mu\nu\rho} \equiv \partial_\mu B_{\nu\rho} + \partial_\nu B_{\rho\mu} + \partial_\rho B_{\mu\nu}, \tag{16.6}$$

as in Problem 10.6. In the last term on the right-hand side of (16.4) we have introduced a dimensionful constant κ needed to make this term dimensionless ($[\kappa^2] = M^{6-D}$). In closed string theory the constant κ is a calculable function of the string coupling and α'. The antisymmetrization is responsible for the factor of $1/2$ which multiplies the second term on the right-hand side of (16.4). We have antisymmetrized the factor multiplying $B_{\mu\nu}$ because it is natural to do so: since $B_{\mu\nu}$ is antisymmetric, any symmetric part of the factor does not contribute to the product. Note the hybrid nature of the action (16.4): part of it is an integral over the string world-sheet, and part of it is an integral over all of spacetime.

In order to appreciate the nature of string charge, we reconsider the Maxwell equations (3.34), where the electric current appears as a source of the electromagnetic field:

$$\frac{\partial F^{\mu\nu}}{\partial x^\nu} = j^\mu. \tag{16.7}$$

Here the electric charge density is j^0. A static particle gives rise to electric charge, but zero electric current. The particle is a source of the Maxwell field, while the string is a source of the $B_{\mu\nu}$ field. There is an equation of motion for the $B_{\mu\nu}$ field, analogous to (16.7). We obtain this equation by calculating the variation of the action (16.4) under a variation $\delta B_{\mu\nu}(x)$. The variation of the last term in the action was calculated in Problem 10.6:

$$\delta\left[-\frac{1}{6\kappa^2}\int d^D x\, H_{\mu\nu\rho}H^{\mu\nu\rho}\right] = \frac{1}{\kappa^2}\int d^D x\, \delta B_{\mu\nu}(x)\,\frac{\partial H^{\mu\nu\rho}}{\partial x^\rho}. \tag{16.8}$$

To vary the second term in S we must vary $B_{\mu\nu}(x)$, but in this term the field is evaluated on the string world-sheet. The field $B_{\mu\nu}(X)$ can be rewritten as an integral over all of spacetime of $B_{\mu\nu}(x)$ times a delta function which localizes the field to the world-sheet:

$$B_{\mu\nu}\left(X(\tau,\sigma)\right) = \int d^D x\ \delta^D\left(x - X(\tau,\sigma)\right) B_{\mu\nu}(x). \tag{16.9}$$

With this identity, the second term in S is rewritten as

$$-\int d^D x B_{\mu\nu}(x)\,\frac{1}{2}\int d\tau d\sigma \delta^D\left(x - X(\tau,\sigma)\right)\frac{\partial X^{[\mu}}{\partial\tau}\frac{\partial X^{\nu]}}{\partial\sigma} \equiv -\int d^D x B_{\mu\nu}(x) j^{\mu\nu}(x), \tag{16.10}$$

where we have introduced the symbol $j^{\mu\nu}$ with value

$$j^{\mu\nu}(x) = \frac{1}{2}\int d\tau d\sigma\, \delta^D\left(x - X(\tau,\sigma)\right)\left(\frac{\partial X^\mu}{\partial\tau}\frac{\partial X^\nu}{\partial\sigma} - \frac{\partial X^\nu}{\partial\tau}\frac{\partial X^\mu}{\partial\sigma}\right). \tag{16.11}$$

It is noteworthy that $j^{\mu\nu}$ is only supported (i.e. it does not vanish) on spacetime points that belong to the string world-sheet. Indeed, if x is not on the world-sheet, then the argument of the delta function is never zero and the integral vanishes. The object $j^{\mu\nu}$ will play the role of a current. By construction, it is antisymmetric under the exchange of its indices:

$$j^{\mu\nu} = -j^{\nu\mu}. \tag{16.12}$$

We have now done all the work needed to find the equation of motion for $B_{\mu\nu}$. Combining equations (16.8) and (16.10), the total variation of the action S is

$$\delta S = \int d^D x\, \delta B_{\mu\nu}(x) \left(\frac{1}{\kappa^2} \frac{\partial H^{\mu\nu\rho}}{\partial x^\rho} - j^{\mu\nu} \right). \tag{16.13}$$

If this variation is to vanish for arbitrary but antisymmetric $\delta B_{\mu\nu}$, the antisymmetric part of the factor multiplying $\delta B_{\mu\nu}$ must vanish (Problem 16.2). Since the entire term in parentheses is antisymmetric, it must vanish:

$$\frac{1}{\kappa^2} \frac{\partial H^{\mu\nu\rho}}{\partial x^\rho} = j^{\mu\nu}. \tag{16.14}$$

Quick calculation 16.1 To test your understanding of antisymmetric variations, consider indices $i, j = 1, 2$ that run over two values and arbitrary antisymmetric variations δB_{ij} such that $\delta B_{ij} G^{ij} = 0$. Show explicitly that the only condition you get is $G^{ij} - G^{ji} = 0$.

The similarity between (16.14) and (16.7) is quite remarkable. It suggests that $j^{\mu\nu}$ is some kind of conserved current. The vector j^μ on the right-hand side of (16.7) is a conserved current because

$$\frac{\partial j^\mu}{\partial x^\mu} = \frac{\partial^2 F^{\mu\nu}}{\partial x^\mu \partial x^\nu} = 0, \tag{16.15}$$

on account of the antisymmetry of $F^{\mu\nu}$ and the exchange symmetry of partial derivatives. In a similar fashion, equation (16.14) gives

$$\frac{\partial j^{\mu\nu}}{\partial x^\mu} = \frac{1}{\kappa^2} \frac{\partial^2 H^{\mu\nu\rho}}{\partial x^\mu \partial x^\rho} = 0. \tag{16.16}$$

The μ index in $j^{\mu\nu}$ is tied to the conservation equation, but the ν index is free. The tensor $j^{\mu\nu}$ can thus be viewed as a set of currents labeled by the index ν. For each fixed ν, the current components are given by the various values of μ. Since the zeroth component of a current is a charge density, we have several charge densities $j^{0\nu}$. More precisely, since $j^{00} = 0$ (16.12), the nonvanishing charge densities are j^{0k}, with k running over spatial values. Therefore the charge densities of a string define a spatial vector:

Kalb–Ramond charge density is a vector $\vec{j}^0$ with components j^{0k}. (16.17)

We will soon prove that the charge density vector is tangent to the string. Consider equation (16.16) for $\nu = 0$:

$$\frac{\partial j^{\mu 0}}{\partial x^\mu} = -\frac{\partial j^{0k}}{\partial x^k} = 0. \tag{16.18}$$

This is the statement that the string charge density is a divergenceless vector:

$$\nabla \cdot \vec{j}^0 = 0. \tag{16.19}$$

The vector string charge $\vec{Q}$ is naturally defined as the space integral of the string charge density:

$$\vec{Q} = \int d^d x \, \vec{j}^0. \tag{16.20}$$

To understand string charge more concretely, let us evaluate $j^{\mu\nu}$ in the static gauge $X^0 = \tau$. With this condition, the delta function in equation (16.11) takes the form

$$\delta\left(x^0 - X^0(\tau,\sigma)\right)\delta\left(\vec{x} - \vec{X}(\tau,\sigma)\right) = \delta(t-\tau)\;\delta\left(\vec{x} - \vec{X}(\tau,\sigma)\right), \tag{16.21}$$

and we can perform the τ integral to find

$$j^{\mu\nu}(\vec{x},t) = \frac{1}{2}\int d\sigma\,\delta\left(\vec{x} - \vec{X}(t,\sigma)\right)\left[\frac{\partial X^\mu}{\partial t}\frac{\partial X^\nu}{\partial \sigma} - \frac{\partial X^\nu}{\partial t}\frac{\partial X^\mu}{\partial \sigma}\right](t,\sigma). \tag{16.22}$$

Clearly, at any fixed time t_0, the current $j^{\mu\nu}$ is supported on the string – the set of points $\vec{X}(t_0,\sigma)$. For j^{0k} the second term in (16.22) does not contribute on account of $X^0 = t$, and we find

$$\vec{j}^0(\vec{x},t) = \frac{1}{2}\int d\sigma\,\delta\left(\vec{x} - \vec{X}(t,\sigma)\right)\vec{X}'(t,\sigma). \tag{16.23}$$

The $\vec{X}'$ factor on the right-hand side of this equation tells us that at every point on the string the string charge density $\vec{j}^0$ is tangent to the string. It points, in fact, along the tangent defined by increasing σ. Since the orientation of a string is said to be the direction of increasing σ, the charge density vector lies along the orientation of the string!

This might seem puzzling. We have emphasized that the reparameterization invariance of the string action means that a change of parameterization cannot change the physics. Changing the direction of increasing σ is a reparameterization, so how can this change the string charge density? While the Nambu–Goto action is invariant under any reparameterization, the coupling (16.3) of the string to the Kalb–Ramond field is *not*. If we let $\sigma \to \pi - \sigma$ while keeping τ invariant, the measure $d\tau d\sigma$ does not change sign but $X^{\nu\prime}$ does. As a result, (16.3) changes sign. In fact, any reparameterization that changes the orientation of the world-sheet will reverse the sign of this term (Problem 16.1).

Open strings are therefore *oriented* curves. At any fixed time they are fully specified by a curve in space together with an identification of the endpoint that corresponds to $\sigma = 0$ (or, equivalently, the endpoint $\sigma = \pi$). Although closed strings do not have endpoints, they still have an orientation, which is also defined by the direction of increasing σ. The open and closed string theories we examined in previous chapters were theories of oriented open strings and oriented closed strings, respectively. Theories of unoriented strings do exist. These are consistent theories obtained by truncating the state space of (oriented) string theories down to the subspace of states that are invariant under the operation of orientation reversal. We examined these theories in a series of problems beginning with Problems 12.12 and 13.5. The theory of unoriented closed strings has no Kalb–Ramond field in the spectrum. This fits in nicely with our discussion since states of unoriented strings do not carry string charge (where could it point to?).

The integral in (16.23) is easily evaluated for an infinitely long static string stretched along the x^1 axis (a similar configuration was studied in Section 6.7). This string is described by the equations

$$X^1(t,\sigma) = f(\sigma), \quad X^2 = X^3 = \cdots = X^d = 0, \tag{16.24}$$

where $f(\sigma)$ is a function of σ whose range is from $-\infty$ to $+\infty$. The function f must be either a strictly increasing or a strictly decreasing function of σ. We expect this distinction to be significant, as these two alternatives correspond to strings with opposite orientation. Only X^1 has σ dependence, so equation (16.23) implies that the only nonvanishing component of $j^{\mu\nu}$ is j^{01} $(= -j^{10})$:

$$\begin{aligned} j^{01}(\vec{x},t) &= \frac{1}{2}\int d\sigma\, \delta\left(x^1 - X^1(\tau,\sigma)\right)\delta(x^2)\delta(x^3)\ldots\delta(x^d)\, f'(\sigma) \\ &= \frac{1}{2}\,\delta(x^2)\delta(x^3)\ldots\delta(x^d)\int_{-\infty}^{\infty} d\sigma\, \delta(x^1 - f(\sigma))f'(\sigma). \end{aligned} \tag{16.25}$$

Letting $\sigma(x^1)$ denote the unique solution of $x^1 - f(\sigma) = 0$, a familiar property of delta functions gives

$$\int_{-\infty}^{\infty} d\sigma\,\delta(x^1 - f(\sigma))f'(\sigma) = \frac{f'(\sigma(x^1))}{|f'(\sigma(x^1))|} = \mathrm{sgn}(f'(\sigma(x^1))), \tag{16.26}$$

where $\mathrm{sgn}(a)$ denotes the sign of a. Since the function f is strictly increasing or strictly decreasing, this sign is either positive or negative for all x^1. Thus, back to $j^{01}(\vec{x},t)$,

$$j^{01}(x^1,\ldots,x^d;t) = \tfrac{1}{2}\,\mathrm{sgn}(f')\;\delta(x^2)\ldots\delta(x^d) = \tfrac{1}{2}\,\mathrm{sgn}(f')\delta(\vec{x}_\perp), \tag{16.27}$$

where $\vec{x}_\perp$ is the vector whose components comprise the directions orthogonal to the string. The string charge density is localized on the string, and we see explicitly the orientation dependence in the sign of f'. For an arbitrary static string the spatial string coordinates X^k are time independent. As a result, equation (16.22) implies that

$$j^{ik} = 0, \quad \text{for a static string.} \tag{16.28}$$

For a static string only the string charge densities j^{0k} are nonvanishing.

Before concluding this section, let us briefly discuss the issue of background fields. We have argued here that the string action must be supplemented by the coupling (16.3). You may ask: were we wrong in our earlier quantization of the string, where we did not consider this extra term in the string action? No, our quantization was valid for zero *background* Kalb–Ramond field. The field $B_{\mu\nu}$ in (16.3) is a called a background Kalb–Ramond field. A Kalb–Ramond background is a $B_{\mu\nu}$ field that satisfies its classical equations of motion. The logical process that led us to consider backgrounds ran as follows. We quantized the closed string and discovered Kalb–Ramond particle states. From these quantum states, we deduced the existence of $B_{\mu\nu}$ fields and derived their (linearized) equations of motion. By postulating that backgrounds exist, we are implying that there are nontrivial $B_{\mu\nu}$ fields that satisfy their full equations of motion. One speaks of background electromagnetic fields,

referring to $\vec{E}$ and $\vec{B}$ fields that satisfy Maxwell's equations. In a related vein, a gravitational background is a spacetime whose metric $g_{\mu\nu}$ satisfies Einstein's equations of general relativity.

Suppose we were handed a $B_{\mu\nu}$ field configuration. How could we decide whether it provides a background? If we had available the full nonlinear field equations, we could simply test whether the field configuration provides a solution. Since we do not have such field equations, a less direct procedure is necessary. We must re-quantize the string, this time using the coupling (16.3) to take into account the effects of the candidate background $B_{\mu\nu}$. If the quantization is successful then we may conclude that the field configuration provides a background. This procedure can be carried out in practice, and physicists have discovered several $B_{\mu\nu}$ backgrounds.

16.2 Visualizing string charge

In classical Maxwell electromagnetism there are many different configurations of charge: point charges, line charges, surface charges, and continuous charge distributions. Since we have seen that the string charge is localized on the string, you may perhaps think that string charge density can be imagined as some Maxwell line charge density on the string. Not really. String *charge density* can be visualized as a Maxwell *current* on the string. Indeed, we saw that string charge density is a spatial vector that points in the direction of the string – this is just what a Maxwell current on the string looks like.

The string charge density can be integrated over space to define the total string charge $\vec{Q}$ (16.20). This charge has some shortcomings: it is infinite for an infinitely long stretched string (see (16.27)), and it vanishes for a contractible closed string (Problem 16.3). It is possible, however, to use the string charge density to count strings. The string number $\mathcal{N}$, to be defined below, counts the number of strings linked by a given space.

The string charge density $\vec{j}^0$ behaves like an electric Maxwell current because of (16.19). Electric charge conservation in electromagnetism requires

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \vec{j} = 0. \tag{16.29}$$

In magnetostatics, the electric charge density ρ is time independent, and, as a result, the electric current density is divergenceless. This means that charge does not accumulate anywhere at any time. Divergenceless currents cannot stop. When they flow on wires either the wires form loops (closed strings for us) or they are infinitely long (infinite strings). We learned that $\nabla \cdot \vec{j}^0 = 0$ vanishes even if we have time-dependent string configurations. Thus electric string charge density is always analogous to an electric current in magnetostatics. String charge conservation requires strings to form closed loops or to be infinitely long.

To elaborate further on the magnetostatic analogy, we examine the Kalb–Ramond fields created by static strings. We will see that the Kalb–Ramond field strength can be encoded

in an effective magnetic field. To simplify matters we work in four-dimensional spacetime. There are two possibilities regarding equation (16.14): either both free indices are space indices or one is a time index and the other is a space index. In the first case we have

$$\frac{\partial H^{ik\rho}}{\partial x^{\rho}} = 0, \tag{16.30}$$

since j^{ik} vanishes for static strings. We satisfy this equation with the following ansatz: all components of H are *time independent* and

$$H^{ijk} = 0. \tag{16.31}$$

The other equation to consider is

$$\frac{\partial H^{0kl}}{\partial x^{l}} = \kappa^2 \, j^{0k}. \tag{16.32}$$

We cast this equation into the form of a Maxwell equation by introducing a vector $\vec{B}_H$ with components $B_{H\,m}$ defined by

$$H^{0kl} = \epsilon^{klm} B_{H\,m}. \tag{16.33}$$

Here ϵ^{ijk} is totally antisymmetric and satisfies $\epsilon^{123} = 1$. The vector $\vec{B}_H$ is called the field strength dual to H. Substituting back into (16.32), we find

$$\epsilon^{klm} \frac{\partial B_{H\,m}}{\partial x^{l}} = \kappa^2 \, j^{0k} \;\longrightarrow\; (\nabla \times \vec{B}_H)_k = \kappa^2 \, j^{0k}. \tag{16.34}$$

At this stage, the relevant components of H have been encoded in a dual "magnetic field," and equation (16.32) has been recast in the form

$$\nabla \times \vec{B}_H = \kappa^2 \, \vec{j}^{0}. \tag{16.35}$$

This is Ampère's equation for the magnetic field of a current $\kappa^2 \vec{j}^{0}$. Note that, given our ansatz, equation (16.35) is equivalent to the original equations for H; if we cannot solve it, there is no solution for H. The consistency condition for (16.35) is familiar. Since the divergence of a curl is zero, the existence of a solution requires (once again) that $\vec{j}^{0}$ be divergenceless. Alternatively, given a closed one-dimensional curve Γ that is the boundary of a two-dimensional surface S, the integral form of equation (16.35) is

$$\frac{1}{\kappa^2} \oint_{\Gamma} \vec{B}_H \cdot d\vec{\ell} = \int_S \vec{j}^{0} \cdot d\vec{a}. \tag{16.36}$$

A curve Γ is said to link a string if the string pierces every surface whose boundary is Γ. If the string ended at some point, the current $\vec{j}^{0}$ would end at that point, as well, leading to a nonvanishing $\nabla \cdot \vec{j}^{0}$ and, consequently, to an inconsistency in (16.35). If the string ended at some point, then for any fixed Γ the left-hand side in (16.36) would be well defined, but the right-hand side would depend on the choice of surface S. This is also an inconsistency.

Equation (16.36) naturally leads to the definition of the string number $\mathcal{N}$ announced at the beginning of this section. The string number $\mathcal{N}$ associated with a curve Γ is defined as

$$\tfrac{1}{2}\mathcal{N} \equiv \frac{1}{\kappa^2} \int_{\Gamma} \vec{B}_H \cdot \vec{d\ell} = \int_S \vec{j}^{0} \cdot d\vec{a}. \tag{16.37}$$

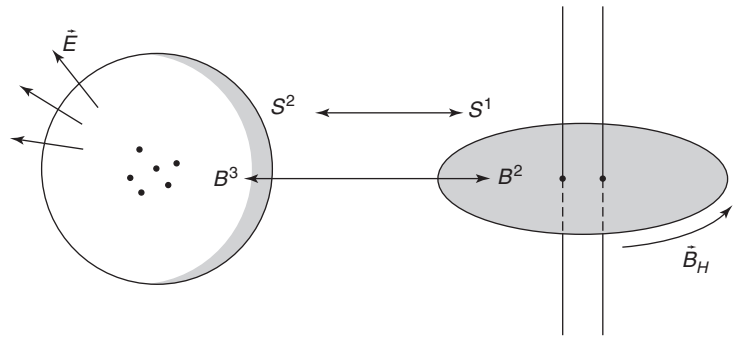

Fig. 16.1 **Comparing the computations of Maxwell charge and of string number in a world with three spatial dimensions. The two-sphere which encloses the Maxwell charge is analogous to the circle which links the strings.**

We expect $\mathcal{N}$ to give the number of strings linked by the curve Γ. Let us calculate, as an illustration, the value of $\mathcal{N}$ for the string stretched along the x^1 axis that we considered in the previous section. We assume, however, that there are only three spatial dimensions, so that the results of the present section apply. Choosing the orientation so that $f'(\sigma) > 0$, equation (16.27) gives

$$j^{01} = \tfrac{1}{2}\delta(y)\delta(z). \tag{16.38}$$

Consider now a closed curve Γ linking the string and lying in a plane of constant x. Assume that on this plane the curve encloses a surface S whose oriented normal points in the positive x direction. Since both the area vector and $\vec{j}^0$ point in the x direction, we find

$$\tfrac{1}{2}\mathcal{N} = \int_S \vec{j}^0 \cdot d\vec{a} = \int_S j^{01} dydz = \tfrac{1}{2}\int_S \delta(y)\delta(z)\; dydz = \tfrac{1}{2}. \tag{16.39}$$

As expected, we got $\mathcal{N} = 1$. In general, $\mathcal{N} = N$, where N is the number of strings linked by the chosen curve. The orientation matters: if the curve links two strings with opposite orientation, then their individual contributions to $\mathcal{N}$ will cancel. The $\vec{B}_H$ field in the above example can be calculated easily. It is also possible to write an explicit expression for the antisymmetric tensor field $B_{\mu\nu}$ (Problem 16.4).

Let us compare with electromagnetism. For a localized Maxwell charge distribution, the charge is calculated by integrating the electric charge density over a three-ball B^3 enclosed by a suitable two-sphere S^2 that encloses the charges (see Figure 16.1). A set of parallel, infinite strings is surrounded – but not enclosed – by a suitable circle S^1. This is the natural analog: in the same way as electric charges do not touch the surface S^2 that encloses them, strings do not touch the "surface" S^1 that links them. You cannot remove a Maxwell charge without puncturing the two-sphere, nor can you remove a string without breaking the circle. The computation analogous to the volume integral of Maxwell charge density, gives the number of strings linked by an S^1 as an integral of the local flux of string charge density over a two-ball B^2 (a disk) whose boundary is the S^1. Finally, in Maxwell theory the

charge can also be computed as a flux integral of the electric field over the surface S^2 which encloses the charges. The string number is computed analogously as an integral of the Kalb–Ramond dual field strength $\vec{B}_H$ along the curve which links the strings.

Quick calculation 16.2 A string lies along the x^1 axis in a world with four spatial dimensions x^1, x^2, x^3, and x^4. Write a couple of equations that define a sphere that links the string.

16.3 Strings ending on D-branes

We learned in Section 15.2 that there is a Maxwell field living on the world-volume of every D-brane. Indeed, photon states arise from the quantization of open strings whose endpoints lie on the D-brane. The quantization of closed strings in Section 13.3 revealed states that arise from a Kalb–Ramond field $B_{\mu\nu}$ living over all spacetime. We have seen that the string couples electrically to the $B_{\mu\nu}$ field. There is therefore an obvious question: if D-branes have Maxwell fields, is there any object that carries electric charge for these fields? This puzzle is related to another one: what happens to the string charge density – which as we learned can be visualized as a current – when a string ends on a D-brane? Does string charge conservation fail to hold?

Puzzles with charge conservation have led to interesting insights in the past. It led, for example, to the recognition of the displacement current in time-dependent electromagnetic processes. In string theory, the solution to the puzzle involves the realization that the ends of the open string behave as electric point charges! They are charged under the Maxwell field that lives on the D-brane where the string ends. Moreover, the electric field lines of those point charges carry string charge. The interplay between string charge and electric charge, and between the associated Kalb–Ramond and Maxwell fields, results in string charge conservation.

Current conservation is intimately related to gauge invariance. In electromagnetism, the coupling of the gauge field to a current is a term in the action which takes the form

$$S_{\text{coup}} = \int d^D x A_\mu(x)\, j^\mu(x). \tag{16.40}$$

In equation (16.2), for example, the coupling term is the middle term on the right-hand side. The gauge transformations are

$$\delta A_\mu(x) = \partial_\mu \epsilon, \tag{16.41}$$

and the field strength $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is gauge invariant: $\delta F_{\mu\nu} = 0$. The first and last terms on the right-hand side of (16.2) are manifestly gauge invariant. This is generic, terms in the action other than (16.40) are gauge invariant by themselves. The gauge invariance of the action then requires the gauge invariance $\delta S_{\text{coup}} = 0$ of the coupling (16.40). Assuming that the current j^μ is itself gauge invariant,

$$\delta S_{\text{coup}} = \int d^D x \ (\partial_\mu \epsilon)\, j^\mu(x) = -\int d^D x \ \epsilon\, \partial_\mu j^\mu(x), \tag{16.42}$$

where we integrated by parts and set the boundary terms to zero by assuming that the parameter ϵ vanishes sufficiently rapidly at infinity. We now see that current conservation ($\partial_\mu j^\mu = 0$) implies gauge invariance ($\delta S_{\text{coup}} = 0$).

Similar ideas hold for the coupling of the Kalb–Ramond field $B_{\mu\nu}$. The gauge transformations of $B_{\mu\nu}$ were given in Problem 10.6:

$$\delta B_{\mu\nu} = \partial_\mu \Lambda_\nu - \partial_\nu \Lambda_\mu. \tag{16.43}$$

The totally antisymmetric field strength $H_{\mu\nu\rho}$ (16.6) is invariant under these gauge transformations. As indicated on the right-hand side of (16.10), the coupling of $B_{\mu\nu}$ to a current $j^{\mu\nu} (= -j^{\nu\mu})$ is of the general form

$$-\int d^D x B_{\mu\nu}(x) j^{\mu\nu}(x). \tag{16.44}$$

Quick calculation 16.3 Prove that the coupling term (16.44) is invariant under the gauge transformations (16.43) if $j^{\mu\nu}$ is a conserved current.

The above results indicate that we can investigate potential failures of current conservation by focusing on the gauge invariance properties of the actions. Let us therefore reconsider the term in the action (16.4) that couples the string to the $B_{\mu\nu}$ field:

$$S_B = -\frac{1}{2}\int d\tau d\sigma \epsilon^{\alpha\beta} \partial_\alpha X^\mu \partial_\beta X^\nu B_{\mu\nu}(X(\tau,\sigma)). \tag{16.45}$$

Here we have introduced two-dimensional indices, $\alpha, \beta = 0, 1$, as well as $\partial_0 = \partial/\partial\tau$ and $\partial_1 = \partial/\partial\sigma$. Also, $\epsilon^{\alpha\beta}$ is totally antisymmetric with $\epsilon^{01} = 1$. Since the gauge invariance of S_B is a little subtle, we will study a simpler case first. We will check the gauge invariance of the term that couples a point particle to the Maxwell field:

$$q\int A_\mu(x) dx^\mu. \tag{16.46}$$

Why is this invariant under (16.41)? Using a parameter τ that ranges from $-\infty$ to $+\infty$, we see that the variation is proportional to

$$\begin{aligned}\int_{-\infty}^{\infty} d\tau\, \delta A_\mu(x(\tau)) \frac{dx^\mu}{d\tau} &= \int_{-\infty}^{\infty} d\tau \frac{\partial \epsilon(x(\tau))}{\partial x^\mu}\frac{dx^\mu}{d\tau} = \int_{-\infty}^{\infty} d\tau \frac{d\epsilon(x(\tau))}{d\tau} \\ &= \epsilon(x(\tau=\infty)) - \epsilon(x(\tau=-\infty)).\end{aligned} \tag{16.47}$$

Since τ parameterizes time, $t(\tau \to \pm\infty) = \pm\infty$. Gauge invariance then follows if we assume that the gauge parameter vanishes in the infinite past and in the infinite future: $\epsilon(t = \pm\infty, \vec{x}) = 0$.

Let us now return to our problem, the gauge invariance of the action (16.45). Since the arguments of $B_{\mu\nu}$ are the string coordinates, the gauge transformations take the form

$$\delta B_{\mu\nu}(X) = \frac{\partial \Lambda_\nu}{\partial X^\mu} - \frac{\partial \Lambda_\mu}{\partial X^\nu}, \tag{16.48}$$

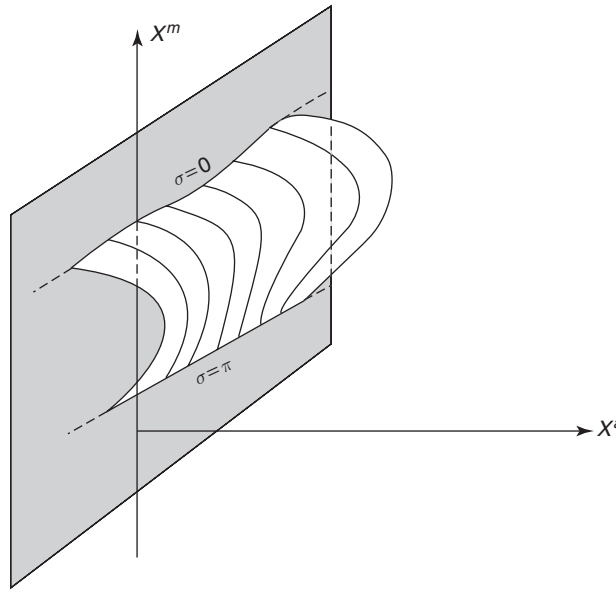

Fig. 16.2 A D-brane and the world-sheet of an open string. The world-sheet boundaries lie on the D-brane: they are the world-lines of the open string endpoints $\sigma = 0$ and $\sigma = \pi$. X^m and X^a are string coordinates along the brane and normal to the brane, respectively.

where the arguments of Λ are also the string coordinates $X(\tau, \sigma)$. The terms multiplying $B_{\mu\nu}$ in (16.45) are antisymmetric in μ and ν (check this!). As a result, each term in (16.48) gives the same contribution to the variation:

$$\delta S_B = -\int d\tau d\sigma\, \epsilon^{\alpha\beta}\, \frac{\partial \Lambda_\nu}{\partial X^\mu}\, \partial_\alpha X^\mu \partial_\beta X^\nu = -\int d\tau d\sigma \epsilon^{\alpha\beta}\, \partial_\alpha \Lambda_\nu\, \partial_\beta X^\nu. \tag{16.49}$$

Writing out the various terms,

$$\begin{aligned} \delta S_B &= -\int d\tau d\sigma \left(\partial_\tau \Lambda_\nu\ \partial_\sigma X^\nu - \partial_\sigma \Lambda_\nu\ \partial_\tau X^\nu\right) \\ &= -\int d\tau d\sigma \Big(\partial_\tau (\Lambda_\nu \partial_\sigma X^\nu) - \partial_\sigma (\Lambda_\nu \partial_\tau X^\nu)\Big). \end{aligned} \tag{16.50}$$

Note that we have two total derivatives. The ∂_τ term gives no contribution since we can assume that Λ vanishes at the endpoints of time. If the string under consideration is closed, then there is no boundary in σ and the ∂_σ term gives no contribution, either. This shows the gauge invariance of S_B for closed strings.

For an open string, however, the ∂_σ term in (16.50) gives rise to boundary contributions that do not vanish. The open string world-sheet has boundaries, which appear as lines on the world-volume of a D-brane (Figure 16.2). Let us now calculate δS_B for open strings. We will call the string coordinates along the brane X^m and the string coordinates normal to the brane X^a:

$$X^\mu = (X^m, X^a), \quad \mu = (m, a). \tag{16.51}$$

If the D-brane is a Dp-brane, then $m = 0, 1, \ldots, p$. We drop the ∂_τ term in (16.50), as before, and focus on

$$\delta S_B = \int d\tau d\sigma\, \partial_\sigma(\Lambda_\nu \partial_\tau X^\nu) = \int d\tau \Big[\, \Lambda_m \partial_\tau X^m + \Lambda_a \partial_\tau X^a \Big]_{\sigma=0}^{\sigma=\pi}. \tag{16.52}$$

Since the X^a are DD coordinates, $\partial_\tau X^a = 0$ at both endpoints, and the second term above gives no contribution. As a result,

$$\delta S_B = \int d\tau\, \Lambda_m \partial_\tau X^m \Big|_{\sigma=\pi} - \int d\tau\, \Lambda_m \partial_\tau X^m \Big|_{\sigma=0}. \tag{16.53}$$

Gauge invariance has failed because of these two boundary terms. This demonstrates that string charge conservation fails at the endpoints of an open string. We must restore gauge invariance. As we have already said, this requires coupling the Maxwell fields on the brane to the ends of the string.

So let us add to the string action a couple of terms that give electric charge to the string endpoints:

$$S = S_B + \int d\tau\, A_m(X) \frac{dX^m}{d\tau}\Big|_{\sigma=\pi} - \int d\tau\, A_m(X) \frac{dX^m}{d\tau}\Big|_{\sigma=0}. \tag{16.54}$$

Since the terms above have opposite signs, the string endpoints are oppositely charged. As a convention, we have chosen the string to begin at the negatively charged endpoint and to end at the positively charged endpoint. We have also set $q = \pm 1$ for the endpoint charges, and we will keep this convention throughout. The physical strength of charges can only be determined if we know the normalization of the F^2 terms on the D-brane. This normalization is fixed by the constant κ_0^2 in (16.2). The F^2 terms, together with the couplings in (16.54), determine how the string endpoints create electromagnetic fields. The normalization of the F^2 terms on the D-brane involve the string coupling and α'. They will be determined in Section 20.3.

More briefly, and in the notation of (16.53), we rewrite (16.54) as

$$S = S_B + \int d\tau\, A_m\, \partial_\tau X^m \Big|_{\sigma=\pi} - \int d\tau\, A_m \partial_\tau X^m \Big|_{\sigma=0}. \tag{16.55}$$

How can we use these terms to restore gauge invariance? By letting the Maxwell field vary under the gauge transformation of the $B_{\mu\nu}$ field! This is a little strange and surprising, but without an interplay between the two types of fields we could not fix our problem of gauge invariance.

So we postulate that, whenever we vary $B_{\mu\nu}$ with a gauge parameter $\Lambda_\mu = (\Lambda_m, \Lambda_a)$, we must also vary the Maxwell field A_m on the D-brane:

$$\begin{aligned} \delta B_{\mu\nu} &= \partial_\mu \Lambda_\nu - \partial_\nu \Lambda_\mu, \\ \delta A_m &= -\Lambda_m. \end{aligned} \tag{16.56}$$

If we vary A_m in this way, then the variation of the last two terms in (16.55) cancels the variations found in (16.53), thus restoring gauge invariance.

Letting A vary as in (16.56) solves the problem at hand, but it raises some interesting questions. Besides the entire string action, we also want the Maxwell action to be gauge invariant. Since this action is proportional to F^2, we ask: is F_{mn} gauge invariant? It is not! Indeed,

$$\delta F_{mn} = \partial_m \delta A_n - \partial_n \delta A_m = -\partial_m \Lambda_n + \partial_n \Lambda_m = -\delta B_{mn}, \tag{16.57}$$

where in the last step we recognized that the variation coincides with the gauge transformation of B_{mn}. This is significant, because it follows that the fully gauge invariant combination is

$$\delta(F_{mn} + B_{mn}) = 0. \tag{16.58}$$

We call the new invariant quantity $\mathcal{F}_{mn}$:

$$\mathcal{F}_{mn} \equiv F_{mn} + B_{mn}, \quad \delta \mathcal{F}_{mn} = 0. \tag{16.59}$$

On the D-brane, $\mathcal{F}_{mn}$ is the physically significant field strength. The familiar field strength F_{mn} is not fully physical because it is not gauge invariant. Maxwell's equations will be modified by replacing F by $\mathcal{F}$. In many circumstances this will be a small modification, and for zero B, $\mathcal{F}$ equals F. The interplay between these fields helps us to understand intuitively the fate of the string charge when a string ends on a D-brane. We turn to this issue now.

We have seen that string charge density can be thought of as a kind of current flowing down the string. Suppose we have a string ending on a D-brane, as shown in Figure 16.3. The current cannot stop flowing at the string endpoint, so it must flow out *into* the D-brane. How can it do so? We know that the string endpoint is charged, so electric field lines emerge from it spreading out inside the D-brane. The field lines cannot go into the ambient space since the Maxwell field only lives on the D-brane. We will see that, in fact, the electric field lines carry the string charge!

As equation (16.10) indicates, string charge density j^{0k} is, by definition, the quantity that couples to B_{0k}. Whatever couples to B_{0k} appears on the right-hand side of (16.14) as a contribution to j^{0k}. On the D-brane, a Lagrangian density proportional to $-\frac{1}{4}\mathcal{F}^{mn}\mathcal{F}_{mn}$, is the gauge invariant generalization of the Maxwell Lagrangian density. Expanding it out,

$$-\frac{1}{4}\mathcal{F}^{mn}\mathcal{F}_{mn} = -\frac{1}{4}B^{mn}B_{mn} - \frac{1}{4}F^{mn}F_{mn} - \frac{1}{2}F^{mn}B_{mn}. \tag{16.60}$$

The last term above is particularly interesting. We can expand it further as

$$-\frac{1}{2}F^{mn}B_{mn} = -F^{0k}B_{0k} + \cdots. \tag{16.61}$$

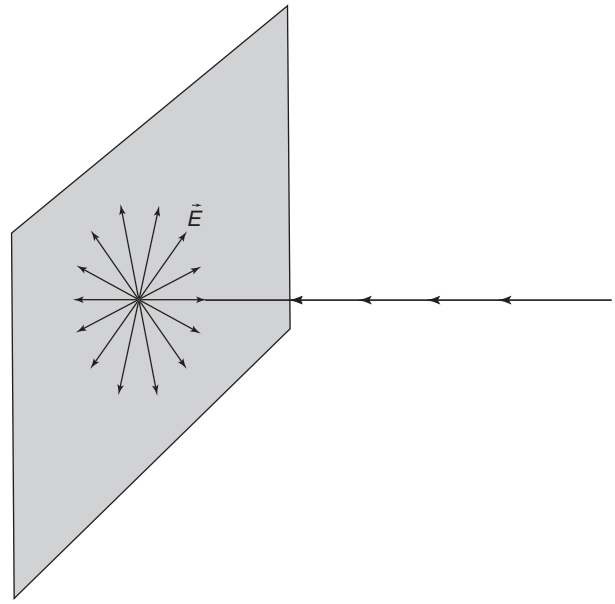

Fig. 16.3 **A string ending on a D-brane. The string charge density carried by the string can be viewed as a current flowing down the string. The current is carried on the D-brane by the electric field lines.**

The complete action for the D-brane and the string will include this $F^{0k}B_{0k}$ term. Anything that couples to B_{0k} carries string charge, so F^{0k} represents string charge on the brane. But $F^{0k} = E_k$ is the electric field. Therefore, electric field lines on the D-brane carry string charge.

16.4 D-brane charges

We have learned that a string carries electric charge for the Kalb–Ramond field of closed string theory. It is natural to wonder whether there are other extended objects in string theory that carry charge, as well. In addition to strings, the only extended objects we have encountered so far are Dp-branes, with various values of p. Do they carry charge?

Point particles have one-dimensional world-lines and carry electric charge if they couple to a one-index massless gauge field. Strings have two-dimensional world-sheets and carry electric charge if they couple to the Kalb–Ramond gauge field, a massless, two-index antisymmetric tensor field. A Dp-brane has a $(p+1)$-dimensional world-volume and is said to be electrically charged if it couples to a massless antisymmetric tensor field with $(p+1)$ indices. The world-volume of the Dp-brane is parameterized by τ and the set of coordinates $\sigma^1, \sigma^2, \ldots, \sigma^p$. The spacetime coordinates that describe the position of the brane are $X^\mu(\tau, \sigma^1, \ldots, \sigma^p)$, with $\mu = 0, 1, \ldots, d$, and the antisymmetric tensor field is denoted by

$A_{\mu\mu_1\ldots\mu_p}(x)$. The required coupling is a generalization of (16.3):

$$S_p = -\int d\tau d\sigma_1 \ldots d\sigma_p \, \frac{\partial X^\mu}{\partial \tau}\frac{\partial X^{\mu_1}}{\partial \sigma^1}\cdots\frac{\partial X^{\mu_p}}{\partial \sigma^p} \, A_{\mu\mu_1\cdots\mu_p}\left(X(\tau,\sigma^1,\ldots,\sigma^p)\right). \tag{16.62}$$

In natural units S_p is dimensionless, so the antisymmetric tensor field has natural units $[\,A_{\mu\mu_1\ldots\mu_p}] = M^{p+1} = L^{-(p+1)}$.

In bosonic closed string theory, the Kalb–Ramond field is the only massless antisymmetric tensor field. This field is sourced by strings, as we have seen. In the absence of additional massless antisymmetric tensors, the bosonic string Dp-branes cannot be charged. On the other hand, type IIA and type IIB closed superstring theories have additional antisymmetric tensors in the Ramond–Ramond sector. These were listed in equations (14.90) and (14.91):

$$\begin{aligned} &\text{IIA}: \; A_\mu, \;\; A_{\mu\nu\rho}, \\ &\text{IIB}: \; A, \;\; A_{\mu\nu}, \;\; A_{\mu\nu\rho\sigma}. \end{aligned} \tag{16.63}$$

It turns out that the R–R gauge fields couple electrically to the appropriate D-branes. In type IIA superstring theory, A_μ couples to D0-branes and $A_{\mu\nu\rho}$ couples to D2-branes. In type IIB superstring theory, $A_{\mu\nu}$ couples to D1-branes and $A_{\mu\nu\rho\sigma}$ couples to D3-branes. The field A in IIB theory carries no index, so it does not couple electrically to any conventional D-brane (it couples electrically to an object called a D-instanton). Summarizing, the electrically charged D-branes are

$$\begin{aligned} &\text{IIA}: \; \text{D0}, \;\; \text{D2}, \\ &\text{IIB}: \; \text{D1}, \;\; \text{D3}. \end{aligned} \tag{16.64}$$

Charge and energy conservation together imply that a charged object cannot decay if there are no lighter mass candidate decay products that can carry the charge. The D-branes in (16.64) are in fact stable D-branes, and they cannot decay into open or closed string states. The bosonic D-branes carry no charge and are unstable, as demonstrated by the existence of a tachyon field on their world-volume. It is known that Dp-branes with even p are stable in type IIA theory but unstable in type IIB theory. The D6-branes of type IIA theory, for example, are stable. Additionally, Dp-branes with odd p are stable in type IIB theory but unstable in type IIA theory. The stable D3-branes of type IIB theory are particularly intriguing because their world-volume is a four-dimensional spacetime. All stable D-branes of type II string theory are charged. Nevertheless, the ones that do not appear in the list (16.64) – the D4, D6, and D8 of type IIA theory, and the D5, D7, and D9 of type IIB theory – turn out to carry *magnetic* charge for either the R–R gauge fields in (16.63) or for other subtle R–R states that we have not included in our discussion. We will not study magnetic charge in this book.

The (electric) charge of a Dp-brane has a simple description when p spatial dimensions are curled up into circles and the Dp-brane is wrapped around the resulting compact space. In this case, the p compact space directions lie along the D-brane (they are of type – in the notation of Table 15.1). The other spacetime directions, which define the effective lower-dimensional spacetime, are normal to the brane (they are of type •). The lower-dimensional observer, who only has access to these noncompact directions, sees the brane as a point

particle. Our claim is that this particle is electrically charged under a Maxwell field that originates from the antisymmetric tensor field $A_{\mu\mu_1\ldots\mu_p}$.

Let $x^1, \ldots, x^p$ denote the p compact directions, and let $X^1, \ldots, X^p$ denote the corresponding brane coordinates. If the compact directions are circles of radii $R^1, \ldots, R^p$ and we use parameters $\sigma^k \in [0, 2\pi]$, then

$$X^k(\tau, \sigma^1, \ldots, \sigma^p) = R^k \sigma^k, \quad k = 1, \ldots, p \text{ (not summed)}, \tag{16.65}$$

represents the wrapped Dp-brane. Indeed, when σ^k runs from zero to 2π, the coordinate X^k runs from zero to $2\pi R^k$, thus going once around the kth circle. Let X^m, with m an index for noncompact directions, take the form

$$X^m(\tau, \sigma^1, \ldots, \sigma^p) = x^m(\tau). \tag{16.66}$$

This equation states that the Dp-brane appears as a point particle to the lower-dimensional observer: for any τ, the full Dp-brane is mapped to a single point in the lower-dimensional spacetime.

Since only X^k depends on σ^k, the nonvanishing contributions to (16.62) arise when $\mu_k = k$, for $k = 1, \ldots, p$:

$$S_p = -\int d\tau d\sigma_1 \ldots d\sigma_p \, \frac{\partial X^\mu}{\partial \tau} R^1 R^2 \ldots R^p \, A_{\mu\, 12\ldots p} \Big(X(\tau, \sigma^1, \ldots, \sigma^p) \Big). \tag{16.67}$$

The tensor field $A_{\ldots}$ is fully antisymmetric and all compact indices have been used, so the index μ can only take values over the noncompact directions: $\mu = m$. As a result,

$$S_p = -\int d\tau d\sigma_1 \ldots d\sigma_p \, \frac{dX^m}{d\tau} R^1 R^2 \ldots R^p \, A_{m\, 12\ldots p} \big(X^m(\tau), X^k(\sigma^k) \big). \tag{16.68}$$

Finally, let us restrict our attention to the part of the $A_{\ldots}$ field that is independent of the compact coordinates: $A_{m\, 12\ldots p}(x^m(\tau))$. Equation (16.68) then becomes

$$S_p = -R^1 R^2 \ldots R^p \int d\tau d\sigma_1 \ldots d\sigma_p \, \frac{dx^m}{d\tau} \, A_{m\, 12\ldots p}\,(x(\tau))\,. \tag{16.69}$$

The σ integrals can now be done, giving a factor of $(2\pi)^p$, which, together with the product of the radii, yields the volume $V_p = (2\pi R^1) \ldots (2\pi R^p)$ of the compact space. Noting that, for all intents and purposes, $A_{m\, 12\ldots p}$ is a one-index gauge field, we introduce the gauge field $\bar{A}_m$, defined by

$$\frac{1}{(\alpha')^{p/2}} \, \bar{A}_m(x(\tau)) \equiv A_{m\, 12\ldots p}\,(x(\tau))\,. \tag{16.70}$$

The factor of α' was introduced in order to give the gauge field $\bar{A}_m$ the expected dimension of mass, or inverse length. The field $\bar{A}_m$ is said to be the Maxwell field that arises from the tensor field $A_{\ldots}$ by dimensional reduction. In this process of dimensional reduction we did two things: (1) all indices except one were taken to run over compact dimensions, and (2) we dropped the dependence of the field on the compact dimensions. We will examine dimensional reduction further in Section 17.6. Using (16.70), the value of S_p in (16.69) becomes

$$S_p = -\frac{V_p}{(\alpha')^{p/2}} \int d\tau \, \frac{dx^m}{d\tau} \, \bar{A}_m\,(x(\tau)) = -\frac{V_p}{(\ell_s)^p} \int \bar{A}_m \, dx^m, \tag{16.71}$$

which is recognized as the coupling of a point particle to a Maxwell field $\bar{A}_m$. The Dp-brane appears as a charged point particle. The Maxwell charge Q of the brane is

$$Q = \frac{V_p}{(\ell_s)^p}. \tag{16.72}$$

The charge Q is given by the volume of the brane, measured in units of string length to the pth power. Q is dimensionless, as it should be.

Problems

Problem 16.1 Reparameterization invariance of the string/Kalb–Ramond coupling.

Consider a world-sheet with coordinates (τ, σ) and a reparameterization that leads to coordinates $(\tau'(\tau, \sigma), \sigma'(\tau, \sigma))$. Show that the coupling (16.3) transforms as follows:

$$\int d\tau' d\sigma' \, \frac{\partial X^\mu}{\partial \tau'} \frac{\partial X^\nu}{\partial \sigma'} \, B_{\mu\nu}(X) = \mathrm{sgn}(\gamma) \int d\tau d\sigma \, \frac{\partial X^\mu}{\partial \tau} \frac{\partial X^\nu}{\partial \sigma} \, B_{\mu\nu}(X),$$

where $\mathrm{sgn}(\gamma)$ denotes the sign of γ and

$$\gamma = \frac{\partial \tau}{\partial \tau'} \frac{\partial \sigma}{\partial \sigma'} - \frac{\partial \tau}{\partial \sigma'} \frac{\partial \sigma}{\partial \tau'}.$$

Note that your proof requires the antisymmetry of $B_{\mu\nu}$. If $\mathrm{sgn}(\gamma) = +1$, the reparameterization is orientation preserving. If $\mathrm{sgn}(\gamma) = -1$, the reparameterization is orientation reversing. Give two examples of nontrivial orientation preserving reparameterizations and two examples of nontrivial orientation reversing reparameterizations.

Problem 16.2 Antisymmetric variations and equations of motion.

Let $\delta B_{\mu\nu} = -\delta B_{\nu\mu}$ be an arbitrary antisymmetric variation ($\mu, \nu = 0, 1, \ldots, d$). Show that

$$\delta B_{\mu\nu} G^{\mu\nu} = \sum_{\mu>\nu}^{d} \sum_{\nu=0}^{d} \delta B_{\mu\nu} (G^{\mu\nu} - G^{\nu\mu}).$$

Now show that if $\delta B_{\mu\nu} G^{\mu\nu} = 0$ for all antisymmetric variations $\delta B_{\mu\nu}$ then $G^{\mu\nu} - G^{\nu\mu} = 0$.

Problem 16.3 Properties of the string charge $\vec{Q}$.

(a) Consider a string at some fixed time t_0 and a region $\mathcal{R}$ of space that contains a portion of this string: the string enters the region $\mathcal{R}$ at a point $\vec{x}_i$ and leaves the region $\mathcal{R}$ at a point $\vec{x}_f$ (assume there is no compactification of space). Use (16.23) to calculate the string charge $\vec{Q} = \int_{\mathcal{R}} d^d x \, \vec{j}^0$ contained in $\mathcal{R}$ at time t_0. Use your result to show that the total string charge $\vec{Q}$ associated with a closed string is zero.

(b) A more abstract proof that $\vec{Q}$ is zero for any localized configuration of closed strings requires showing that $\nabla \cdot \vec{j}^0 = 0$ implies $\int d^d x \, \vec{j}^0 = 0$. If you have trouble showing this you may look in your favorite E&M book: the same proof is needed in magnetostatics to demonstrate that the multipole expansion for the magnetic field of a localized current has no monopole term.

(c) Assume now that one space coordinate x is curled up into a circle of radius R, and consider a closed string wrapped around this circle. Calculate the string charge $\vec{Q}$. Explain why the answer is not zero, and compare with the result obtained in (16.72).

Problem 16.4 Kalb–Ramond field of a string.

Following the discussion in Section 16.2, calculate the field $H^{\mu\nu\rho}$ created by a string stretched along the x axis. Find a simple $B^{\mu\nu}$ that gives rise to the field strength H. Interpret your answer in terms of a magnetostatics analog.

Problem 16.5 Explicit checks of current conservation.

In Problem 5.3 you constructed the current vector of a charged point particle:

$$j^\mu(\vec{x}, t) = qc \int d\tau\, \delta^D(x - x(\tau))\, \frac{dx^\mu(\tau)}{d\tau}.$$

Verify directly that this is a conserved current ($\partial_\mu j^\mu = 0$). Now extend this result to the case of a closed string. Verify directly that the current in (16.11) is conserved.

Problem 16.6 Equation of motion for a string in a Kalb–Ramond background.

Consider the string action (6.39) supplemented by the coupling (16.3) to the Kalb–Ramond field. Perform a variation δX^μ, and prove that the equations of motion for the string are

$$\frac{\partial \mathcal{P}^\tau_\mu}{\partial \tau} + \frac{\partial \mathcal{P}^\sigma_\mu}{\partial \sigma} = -H_{\mu\nu\rho} \frac{\partial X^\nu}{\partial \tau} \frac{\partial X^\rho}{\partial \sigma}. \tag{1}$$

Problem 16.7 H-field and a circular closed string.

Assume we have a constant, uniform H field that takes the value $H_{012} = h$, with all other components equal to zero. Assume we also have a circular closed string lying in the (x^1, x^2) plane. The purpose of this problem is to show that the tension of the string and the force on the string due to H can give rise to an equilibrium radius. As we will also see, the equilibrium is unstable.

We will analyze the problem in two ways. First, we use the equation of motion (1) derived in Problem 16.6, working with $X^0 = \tau$ ($c = 1$):

(a) Find simplified forms for $\mathcal{P}^\tau_\mu$ and $\mathcal{P}^\sigma_\mu$ for a static string.
(b) Check that the $\mu = 0$ component of the equation of motion is trivially satisfied. Show that the $\mu = 1$ and $\mu = 2$ components give the same result: for a suitable orientation of the closed string, the radius R of the string is fixed at the value $R = T_0/|h|$.

Second, we evaluate the action using the simplified geometry of the problem. For this, assume that the radius $R(t)$ is time dependent.

(c) Find $B_{\mu\nu}$ fields that give rise to the H field. In fact, you can find a solution where only B_{01} or B_{02}, or both, are nonzero.
(d) Show that the coupling term (16.3) for the string in question is equal to

$$\int dt\, \pi h R^2(t)$$

if the string is oriented counterclockwise in the (x_1, x_2) plane. Explain why this term represents (minus) potential energy.

(e) Consider the full action for this circular string, and use this to compute the energy functional $E(R(t), \dot{R}(t))$ (your analysis of the circular string in Problem 6.7 may save you a little work).

(f) Assume $\dot{R}(t) = 0$, and plot the energy functional $E(R(t))$ for $h > 0$ and $h < 0$. Show that the equilibrium value of the radius coincides with the one obtained before, and explain why this equilibrium value is unstable.

17 T-duality of closed strings

If a spatial dimension is curled up into a circle then closed strings are affected in two ways: their momentum along the circle gets quantized, and new winding states that wrap around the circle arise. The complementary behavior of momentum and winding states, as a function of the radius of the circle, results in a surprising symmetry: in closed string theory, the physics when the circle has radius R is indistinguishable from the physics when the circle has radius α'/R. This equivalence is proven by exhibiting an operator map between the two theories that respects all commutation relations.

17.1 Duality symmetries and Hamiltonians

Duality symmetries are some of the most interesting symmetries in physics. The term "duality" is generally used by physicists to refer to the relationship between two systems that have very different descriptions but identical physics. The main subject of this chapter is one such situation that arises in closed string theory. You may think that a world where one dimension is curled up into a circle of radius R could easily be distinguished from a world in which the circle has radius α'/R (recall that α' has units of length-squared), but in closed string theory these two worlds are indistinguishable for any value of R. There is a duality symmetry that relates them to each other. This symmetry is called T-duality, where the T stands for toroidal. A compactification is called toroidal if the compact space is a torus. With this terminology, a one-dimensional torus is defined to be a circle.

The AdS/CFT correspondence, to be described in Chapter 23, is an example of a duality: a type IIB superstring background and a supersymmetric Yang–Mills theory are in fact physically equivalent systems. In this section we will discuss duality symmetries that can be found in two familiar situations. Our first example is from electromagnetism and the second is from mechanics.

Consider Maxwell's equations in the absence of sources:

$$\begin{aligned} \nabla \cdot \vec{E} = 0\,, \quad \nabla \times \vec{B} &= \frac{1}{c}\frac{\partial \vec{E}}{\partial t}\,, \\ \nabla \cdot \vec{B} = 0\,, \quad \nabla \times \vec{E} &= -\frac{1}{c}\frac{\partial \vec{B}}{\partial t}\,. \end{aligned} \tag{17.1}$$

One immediately notices that these equations remain invariant under the following *duality transformation*:

$$(\vec{E}\,,\,\vec{B}) \;\rightarrow\; (-\vec{B}\,,\,\vec{E})\,. \tag{17.2}$$

This invariance of the equations of motion is called the duality symmetry of electromagnetism. Had we simply looked at the familiar Lagrangian for electromagnetism, we might have missed this symmetry. Indeed, expanding the Lagrangian density (see, for example, Problem 5.6) in terms of electric and magnetic fields, we have

$$\begin{aligned}\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} &= -\frac{1}{4}(2F_{0k}F^{0k} + F_{ij}F^{ij}) \\ &= -\frac{1}{2}(-F_{0k}F_{0k} + \frac{1}{2}F_{ij}F_{ij}) \\ &= \frac{1}{2}\,(E^2 - B^2)\,,\end{aligned} \tag{17.3}$$

where we wrote $E^2 = \vec{E}\cdot\vec{E}$ and $B^2 = \vec{B}\cdot\vec{B}$. The Lagrangian density $\mathcal{L}$ is *not* invariant under the duality transformation (17.2) – it changes sign. Of course, $\mathcal{L}$ and $-\mathcal{L}$, both written in terms of potentials, lead to the same equations of motion.

Lagrangians treat kinetic and potential energies differently: kinetic energy enters with a positive sign while potential energy enters with a negative sign. Any symmetry that exchanges these types of energy fails to leave the Lagrangian invariant. Both potential and kinetic energies enter into the Hamiltonian with the same sign, so duality symmetries are often exhibited using the Hamiltonian. In electromagnetism, the square of the electric field accounts for the kinetic energy since the electric field involves the time derivative of the vector potential (see equation (3.8)). The square of the magnetic field accounts for the potential energy since the magnetic field involves spatial derivatives of the vector potential. The Hamiltonian, or energy functional, is proportional to the volume integral of $(E^2 + B^2)$. The duality transformation (17.2) leaves the Hamiltonian unchanged.

Since the dynamical variables in electromagnetism are the gauge potentials, one may ask: are there transformations of the potentials that induce (17.2)? Not exactly, but close. One can formulate electromagnetism (without sources) using dynamical variables $\vec{E}$ and $\vec{A}$, both divergenceless. Duality transformations are then written as spatially nonlocal transformations of $\vec{E}$ and $\vec{A}$. (A typical spatially nonlocal transformation expresses the output at a point $\vec{x}$ in terms of the values of the input at all points $\vec{x}'$.)

Consider now a second example. The system is a simple harmonic oscillator consisting of a mass m attached to a spring with spring constant k. The Hamiltonian is given by

$$H(m,k) = \frac{p^2}{2m} + \frac{1}{2}kx^2\,, \tag{17.4}$$

where we have included the parameters m and k as arguments of the Hamiltonian. This Hamiltonian leads to oscillatory motion with angular frequency $\omega = \sqrt{k/m}$. The form of ω suggests a symmetry under the following duality transformation:

$$(m\,,\,k\,) \;\longrightarrow\; \Big(\frac{1}{k}\,,\;\frac{1}{m}\Big)\,. \tag{17.5}$$

The Lagrangian for the original oscillator

$$L = \frac{1}{2}m\dot{x}^2 - \frac{1}{2}kx^2\,, \tag{17.6}$$

just as in the electromagnetic case, does not remain invariant under the duality transformation. On the other hand, the equation of motion

$$m\ddot{x} = -kx\,, \tag{17.7}$$

is invariant under the duality. The Hamiltonian associated with the dual parameters is

$$H\Big(\frac{1}{k}\,,\frac{1}{m}\Big) = \frac{1}{2}\,kp^2 + \frac{1}{2m}x^2\,. \tag{17.8}$$

Unlike before, now the Hamiltonian is *not* invariant. To exhibit the connection to the original Hamiltonian $H(m,k)$ we must use canonical transformations. This means changing the canonical variables in such a way that all commutation relations are preserved and therefore the physics remains the same. Consider the canonical transformation K that acts on x and p as follows:

$$K\colon\ x \longrightarrow p\,,\quad p \longrightarrow -x\,. \tag{17.9}$$

This transformation is canonical because all the relevant commutation relations, which in this case is only $[x,p]=i$, are preserved:

$$K\colon [\,x,p\,] \longrightarrow [\,p,-x\,] = -(-i) = i\,. \tag{17.10}$$

Under this canonical transformation the Hamiltonian in (17.8) becomes

$$K\colon H\Big(\frac{1}{k}\,,\frac{1}{m}\Big) \longrightarrow \frac{1}{2}\,k\,(-x)^2 + \frac{1}{2m}\,p^2 = H(m,k)\,. \tag{17.11}$$

The Hamiltonian of the system with dual parameters is canonically equivalent to the original Hamiltonian, so the physics is indeed unchanged by the transformation (17.5).

17.2 Winding closed strings

To explore T-duality of closed string theory, we must first understand what effects there are on closed strings when one spatial dimension has been made into a circle. The closed strings we considered in Chapter 13 were moving in Minkowski space, and they could all be shrunk to zero size continuously. If we have one compact dimension then not all closed strings can be reduced continuously to zero size.

The easiest way to visualize this phenomenon is to imagine a world with only two spatial dimensions, one of which is compact. Such a world can be thought of as the surface of an infinitely long cylinder. Let x be the coordinate that has been made compact via the identification

$$x \sim x + 2\pi R\,, \tag{17.12}$$

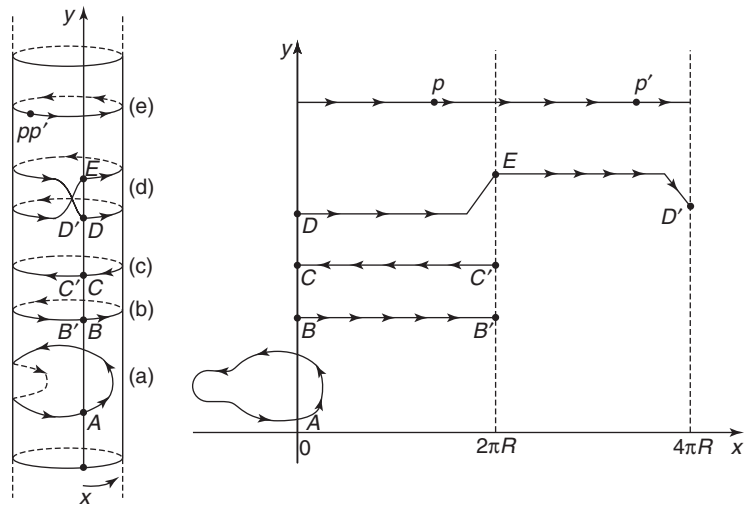

Fig. 17.1 **Left: a collection of closed strings living on the surface of a two-dimensional cylinder. Right: the same set of strings represented on the covering space of the cylinder. Strings with nontrivial winding numbers appear in the covering space as open strings.**

and let y denote the coordinate that extends along the length of the cylinder. This cylinder is shown on the left side of Figure 17.1. On the right side of the figure we show the (x, y) plane from which the cylinder was formed using the identification (17.12). This plane is known as the covering space of the cylinder. As usual, the x coordinate of a string will be denoted by X.

Let us consider what kinds of closed strings can live on the two-dimensional surface of the cylinder. We will also examine, on the right, how these strings appear in the covering space. While every string on the cylinder actually has infinitely many copies in the covering space (recall Section 2.7), to avoid confusion we will show only one copy.

The simplest strings are those that do not wrap around the cylinder. String (a) in Figure 17.1 is such a string. This string, without its copies, is shown in the covering space directly to the right, where it appears as a string that is actually closed. String (a) can be contracted down to a point in a continuous way. This is clear both on the cylinder and in the covering space. We say that string (a) has zero winding number because it does not "wind around" the compact dimension. The string coordinate in the covering space satisfies

$$\text{string (a):} \quad X(\tau, \sigma = 2\pi) - X(\tau, \sigma = 0) = 0\,. \tag{17.13}$$

This is the periodicity condition that we encountered in Chapter 13, where, more generally, we wrote $X(\tau, \sigma + 2\pi) = X(\tau, \sigma)$.

Consider now string (b) in the figure. This string is oriented in the direction of increasing x and wraps once around the cylinder. String (b) cannot be contracted to a point without first cutting it. We say that string (b) has winding number $+1$, since it winds once around the circle in the direction of positive x. The picture of this string in the covering space is interesting. If we mark point B on the cylinder as a reference starting point for the string,

then in the covering space the string looks like an open curve that begins at B and ends at the point B'. We recognize that the string is actually closed because B and B' are identified by (17.12). The closed string is parameterized with $\sigma \in [0, 2\pi]$, where B corresponds to $\sigma = 0$, and B' corresponds to $\sigma = 2\pi$. The string coordinate in the covering space then satisfies

$$\text{string (b):} \quad X(\tau, \sigma = 2\pi) - X(\tau, \sigma = 0) = 2\pi R\,. \tag{17.14}$$

Analogous remarks hold for string (c). This string wraps around the cylinder once in the opposite direction, so it is said to have winding number -1. In the covering space we have

$$\text{string (c):} \quad X(\tau, \sigma = 2\pi) - X(\tau, \sigma = 0) = -2\pi R\,. \tag{17.15}$$

Strings (b) and (c) illustrate a general fact. Strings which wind around the cylinder appear in the covering space as open strings. But not every open string in the covering space represents a wound closed string. Some represent truly open strings on the cylinder. An open string in the covering space represents a closed string that is wound around the cylinder precisely when its endpoints are identified by (17.12). Since string (a) has zero winding, it is a closed string in the covering space.

Consider now a string, such as string (d), that wraps twice around the cylinder. The starting reference point is taken to be D, at which point both $x = 0$ and $\sigma = 0$. After the string has wrapped once it reaches point E on the cylinder. It then winds once more to finally reach point D' with $\sigma = 2\pi$. Points D and D' are the same on the cylinder. In the covering space they are separate, but they are identified by (17.12). This string is said to have winding number $+2$, and it satisfies

$$\text{String (d):} \quad X(\tau, \sigma = 2\pi) - X(\tau, \sigma = 0) = 2\,(2\pi R)\,. \tag{17.16}$$

String (e) is a string with winding number $+2$, just like string (d). But the y coordinate of string (e) is a constant function of σ. Since it has the same winding number as string (d), it also satisfies condition (17.16). Back on the cylinder the windings overlap. In the covering space the string is represented as a straight horizontal line of length $4\pi R$.

A small confusion is possible here. Consider two distinct points p and p' *on* string (e) that are separated by a distance $2\pi R$ along the string. These two points happen to lie on top of the same point on the cylinder, but they are *not* the same point on the wrapped string. In the covering space the points p and p' of the string are separated by a horizontal distance of $2\pi R$. They lie on top of points that are identified, but p and p' are themselves not identified. You can imagine painting one half of the string yellow and the other half green. Point p would belong on the yellow part of the string and point p' would belong on the green part. The piece of string that stretches along the interval $[2\pi R, 4\pi R]$ in the covering space is not a copy of the piece of string that stretches along $[0, 2\pi R]$.

In Figure 17.1 we omitted all duplicate strings from the covering space. Each portion of string that is drawn represents a distinct portion of the string on the cylinder. This was done precisely in order to avoid the above confusion. If, for example, we had included the copies of string (b) in the covering space, then it would be impossible to distinguish between strings (b) and (e) without using colors to draw them differently in the covering space.

All closed strings, however long they may be, and however many times they may wrap, are parameterized with a σ range of 2π. We say that a string has *winding number* m, with m an integer, if it wraps m times around the cylinder in the direction of positive x. In this case, the string coordinate in the covering space satisfies

$$\text{winding number } m: \quad X(\tau, \sigma + 2\pi) = X(\tau, \sigma) + m(2\pi R)\,. \tag{17.17}$$

Strings with different winding numbers cannot be continuously deformed into each other, and therefore the winding number of a closed string is a topological property. Mathematically, winding numbers appear because we have *two* circles: one with coordinate σ and one with coordinate x. The closed strings are mappings from the σ-circle into the x-circle. A mapping of one circle into another is characterized by an integer known as the winding number of the map. For closed strings, this integer is the winding number m.

For purposes that will become clear later, we define the *winding* w in terms of the winding number m and the radius of the space:

$$w \equiv \frac{mR}{\alpha'}\,. \tag{17.18}$$

The winding has units of inverse length, or momentum. In fact, winding will turn out to be a new kind of momentum. Using this definition, equation (17.17) becomes

$$X(\tau, \sigma + 2\pi) = X(\tau, \sigma) + 2\pi\alpha' w\,. \tag{17.19}$$

We are now ready to discuss the mode expansion and quantization of closed strings in the presence of a compact dimension.

17.3 Left movers and right movers

Let us consider now strings that are propagating in a 26-dimensional spacetime that has coordinates $x^0, x^1, \ldots, x^{25}$. Assume that the coordinate x^{25} is curled up into a circle of radius R. Since we are going to use the light-cone gauge (this is possible because the string coordinates X^0 and X^1 are not associated with compact directions), it is convenient to organize the string coordinates as

$$X^+, X^-, \underbrace{X^2, X^3, \ldots, X^{24}}_{X^i}, X^{25}\,. \tag{17.20}$$

X^i denotes transverse light-cone coordinates, but we do not permit the i index to take the value 25. We will always write the coordinate index explicitly, so we can delete the

superscript from X^{25} without risk of confusion. With this understanding, the above set of coordinates will be represented by

$$X^+, X^-, \{X^i\}, \quad \text{and} \quad X \quad \text{with} \quad i = 2, 3, \ldots, 24\,. \tag{17.21}$$

The periodicity condition for X is given by (17.19). Since X satisfies the wave equation, the general solution is

$$X(\tau, \sigma) = X_L(\tau + \sigma) + X_R(\tau - \sigma) = X_L(u) + X_R(v)\,, \tag{17.22}$$

where $u = \tau + \sigma$ and $v = \tau - \sigma$. Applying condition (17.19) we get

$$X_L(u + 2\pi) + X_R(v - 2\pi) = X_L(u) + X_R(v) + 2\pi\alpha' w\,, \tag{17.23}$$

which can be rewritten as

$$X_L(u + 2\pi) - X_L(u) = X_R(v) - X_R(v - 2\pi) + 2\pi\alpha' w\,. \tag{17.24}$$

The last term in the right-hand side was not present when we studied closed strings in Chapter 13 (see (13.13)). Some of the earlier analysis still applies, however. Just as before, the derivatives $X_L'(u)$ and $X_R'(v)$ are periodic functions of their arguments. So the expansions (13.14) hold in the present case, and equations (13.15) hold as well:

$$\begin{aligned} X_L(u) &= \frac{1}{2}x_0^L + \sqrt{\frac{\alpha'}{2}}\bar{\alpha}_0\, u + i\sqrt{\frac{\alpha'}{2}}\sum_{n\neq 0}\frac{\bar{\alpha}_n}{n}e^{-inu}\,, \\ X_R(v) &= \frac{1}{2}x_0^R + \sqrt{\frac{\alpha'}{2}}\alpha_0\, v + i\sqrt{\frac{\alpha'}{2}}\sum_{n\neq 0}\frac{\alpha_n}{n}e^{-inv}\,. \end{aligned} \tag{17.25}$$

The new feature is that $\bar{\alpha}_0$ does not necessarily equal α_0. Instead, equation (17.24) gives

$$2\pi\sqrt{\frac{\alpha'}{2}}\,\bar{\alpha}_0 = 2\pi\sqrt{\frac{\alpha'}{2}}\,\alpha_0 + 2\pi\alpha' w\,, \tag{17.26}$$

so, after cancelling constants,

$$\bar{\alpha}_0 - \alpha_0 = \sqrt{2\alpha'}\, w\,. \tag{17.27}$$

We see that now $\bar{\alpha}_0$ is equal to α_0 if and only if the winding vanishes. We can also calculate the momentum p of the string along the compact direction:

$$p = \frac{1}{2\pi\alpha'}\int_0^{2\pi}(\dot{X}_L + \dot{X}_R)d\sigma = \frac{1}{\sqrt{2\alpha'}}(\alpha_0 + \bar{\alpha}_0)\,, \tag{17.28}$$

where we used equations (17.3) and noted that only the terms that are linear in u and v contribute to the integral. The momentum is proportional to the average of α_0 and $\bar{\alpha}_0$. In summary, we can now write

$$\begin{aligned} p &= \frac{1}{\sqrt{2\alpha'}}\,(\bar{\alpha}_0 + \alpha_0)\,, \\ w &= \frac{1}{\sqrt{2\alpha'}}\,(\bar{\alpha}_0 - \alpha_0)\,. \end{aligned} \tag{17.29}$$

These equations suggest that the winding w is on the same footing as the momentum p. We think of both w and p as momentum operators. For reference, we record the values of the zero modes:

$$\alpha_0 = \sqrt{\frac{\alpha'}{2}}\,(p - w)\,,$$
$$\bar{\alpha}_0 = \sqrt{\frac{\alpha'}{2}}\,(p + w)\,. \tag{17.30}$$

When there was no compactification we had $\alpha_0 = \bar{\alpha}_0$, and this meant that there was only one momentum. Because of this, we were led to expect that only one coordinate zero mode was relevant. Now, however, we have $\alpha_0 \neq \bar{\alpha}_0$, and so there are two different kinds of momenta. There is therefore room for two distinct coordinate zero modes. We rewrite $x_0^L = x_0 + q_0$ and $x_0^R = x_0 - q_0$, thus introducing the average coordinate x_0 and the coordinate difference q_0. These, together with (17.3), allow us to rewrite (17.3) as

$$X_L(\tau+\sigma) = \frac{1}{2}(x_0+q_0) + \frac{\alpha'}{2}(p+w)(\tau+\sigma) + i\sqrt{\frac{\alpha'}{2}}\sum_{n\neq 0}\frac{\bar{\alpha}_n}{n}e^{-in(\tau+\sigma)}\,,$$
$$X_R(\tau-\sigma) = \frac{1}{2}(x_0-q_0) + \frac{\alpha'}{2}(p-w)(\tau-\sigma) + i\sqrt{\frac{\alpha'}{2}}\sum_{n\neq 0}\frac{\alpha_n}{n}e^{-in(\tau-\sigma)}\,. \tag{17.31}$$

The full coordinate $X(\tau,\sigma)$ is obtained by adding the above expressions for X_L and X_R:

$$X(\tau,\sigma) = x_0 + \alpha' p\tau + \alpha' w\sigma + i\sqrt{\frac{\alpha'}{2}}\sum_{n\neq 0}\frac{e^{-in\tau}}{n}(\bar{\alpha}_n e^{-in\sigma} + \alpha_n e^{in\sigma})\,. \tag{17.32}$$

In this expansion the only evidence of a compact dimension is the winding term $\alpha' w\sigma$. The zero mode q_0 is not present here. For the record, we give the standard linear combinations of derivatives of X:

$$\dot{X} + X' = 2X_L{}'(\tau+\sigma) = \sqrt{2\alpha'}\sum_{n\in\mathbb{Z}}\bar{\alpha}_n\, e^{-in(\tau+\sigma)}\,,$$
$$\dot{X} - X' = 2X_R{}'(\tau-\sigma) = \sqrt{2\alpha'}\sum_{n\in\mathbb{Z}}\alpha_n e^{-in(\tau-\sigma)}\,. \tag{17.33}$$

17.4 Quantization and commutation relations

In this section we will derive the commutation relations for the modes of the string coordinate X. We will then discuss the spectrum of the operators p and w.

The starting point is the familiar set of canonical commutators. The commutator between the string coordinate X and the string momentum $\mathcal{P}^\tau$ is taken to be

$$\left[\, X(\tau,\sigma)\,,\ \mathcal{P}^\tau(\tau,\sigma')\right] = i\,\delta(\sigma-\sigma')\,. \tag{17.34}$$

In addition, the commutators that involve two coordinates or two momenta are presumed to vanish. Since the quantization of all the extended coordinates exactly follows the discussion in Chapter 13, we only need to consider the compact direction. But even for this coordinate X the situation is not so different.

First consider the commutators that do not involve x_0. When we computed these commutators in Chapter 13, the key to the calculation was (13.28). This equation was simple to deal with because the linear combinations of derivatives of X^I that appear in it took the compact form recorded in (13.26). These combinations of derivatives of X, recorded in (17.3), presently take the same form except for two minor differences: there is no superscript index, and α_0 and $\bar\alpha_0$ are not the same. It follows that equation (13.29) applies also to the modes of the string coordinate X, and therefore

$$[\bar\alpha_m\,,\bar\alpha_n] = [\alpha_m\,,\alpha_n] = m\,\delta_{m+n,0}\ , \quad [\alpha_m\,,\bar\alpha_n] = 0\,, \tag{17.35}$$

for all integers m and n. Of particular interest are the commutators involving α_0 and $\bar\alpha_0$. We have $[\alpha_0,\bar\alpha_0]=0$, so on account of (17.3) we find

$$[\,p,w\,] = 0\,. \tag{17.36}$$

Since both α_0 and $\bar\alpha_0$ commute with all the oscillators α_n and $\bar\alpha_n$, so too do p and w:

$$[\,p\,,\bar\alpha_n] = [\,p\,,\alpha_n] = [\,w\,,\bar\alpha_n] = [\,w\,,\alpha_n] = 0\,. \tag{17.37}$$

All that remains is to determine the commutation relations of x_0 with the other operators. The strategy is to use

$$[\,X(\tau,\sigma),\ (\dot X \pm X')(\tau,\sigma')] = 2\pi\alpha'\,i\,\delta(\sigma-\sigma')\,, \tag{17.38}$$

which is derived by combining (17.34) and the σ' derivative of $[X(\tau,\sigma),X(\tau,\sigma')]=0$. Glancing at (17.32), we see that the terms including p and w do not contribute, since p and w commute with all the α_n and $\bar\alpha_n$ operators. Integrating over $\sigma\in[0,2\pi]$, we find

$$\left[\, x_0\,,\ (\dot X \pm X')(\tau,\sigma')\,\right] = \alpha'\,i\ . \tag{17.39}$$

It follows from this equation and the expansions (17.3) that x_0 commutes with all the α_n and $\bar\alpha_n$ operators of nonzero label, and that

$$[\,x_0\,,\alpha_0] = [\,x_0\,,\bar\alpha_0] = i\sqrt{\frac{\alpha'}{2}}\ . \tag{17.40}$$

Comparing with (17.3), we see that

$$[\,x_0\,,\,p\,] = i\,, \quad [\,x_0\,,w\,] = 0. \tag{17.41}$$

This completes our analysis of the commutation relations. There is one surprising point worth noticing: every operator that appears in X commutes with the winding w. The most conservative interpretation of this result is to say that w is a constant number. This would

mean that the above quantization is only able to describe the set of closed strings that have some particular fixed winding. Different windings would correspond to different sectors of the full closed string theory. This interpretation treats p and w in quite different ways. A more intriguing interpretation is that w is an operator, just like p is, and that the eigenvalues of w correspond to the various possible windings. This will turn out to be the more natural interpretation, and, as we will see in Section 17.8, it is possible to identify the coordinate zero mode q_0 as the coordinate conjugate to the momentum w.

Because we have compactified the x dimension, the zero mode x_0 is a coordinate that lives on a circle. The momentum operator p along the x direction is the momentum conjugate to x_0, so by a familiar result in quantum mechanics, the possible values of the p-momentum carried by the states are quantized. To derive this quantization condition consider the operator e^{-iap} that translates states along the x direction by an amount a. Since x_0 lives on a circle of radius R, the translation operator that translates by $2\pi R$ has no effect on the states of the theory. Thus $e^{-i\,2\pi Rp}$ behaves like the unit operator. From this we conclude that the states of the theory have momentum along x that is quantized to take the values

$$p = \frac{n}{R}\,, \quad n \in \mathbb{Z}\,. \tag{17.42}$$

It should be noted that x_0 is not a well defined operator: its eigenvalues are ambiguous because of the identification $x_0 \sim x_0 + 2\pi R$. As a result, the commutation relation $[x_0, p] = i$ is in fact a formal statement without a precise meaning. The simplest form of the uncertainty principle, which states that for momentum eigenstates the position uncertainty is infinite, does not apply: in a circle the maximum position uncertainty is $2\pi R$. We can use x_0 to construct the well defined operators

$$e^{i\ell x_0/R}\,, \quad \ell \in \mathbb{Z}\,, \tag{17.43}$$

which are invariant under the shift of x_0 by any multiple of $2\pi R$. We then have

$$e^{-i\ell x_0/R}\, p\, e^{i\ell x_0/R} = p + \frac{\ell}{R}\,. \tag{17.44}$$

This operator equation is a precise statement. It is easily derived using $[x_0, p] = i$. It is in this sense that the naive relation $[x_0, p] = i$ can be used to derive unambiguous results. Similar remarks actually apply to (17.34) since X is not a well defined operator either. The operators in (17.43) have a well defined action on the state space, described in terms of momentum eigenstates $|p\rangle$.

Quick calculation 17.1 Show that $e^{i\ell x_0/R}|p\rangle$ is a state of momentum $p + (\ell/R)$. Note that this momentum is properly quantized.

There is another quantization condition in play. Recalling the periodicity condition (17.17), we demand that

$$X(\tau, \sigma + 2\pi) = X(\tau, \sigma) + m(2\pi R) \tag{17.45}$$

acting on the allowed states of the theory. On account of (17.32), this requires that the operator w has eigenvalues satisfying

$$\alpha' w(2\pi) = m(2\pi R) \longrightarrow w = \frac{mR}{\alpha'}, \quad m \in \mathbb{Z}. \tag{17.46}$$

Thus both p and w have discrete spectra.

There is an interesting double strike mechanism at work here. The identification $x \sim x + 2\pi R$ had two effects, or strikes. First, we lost some states. Without the circle, the momentum operator has a continuous spectrum. After the circle is created, the momentum is quantized, so we lost the states that do not satisfy the quantization condition. Second, we gained some new states. These are the winding states that wrap around the newly created circle. Thus we both lost some states and gained some states! It is interesting to note that for particle states, as opposed to string states, we only lose states when we turn a line into a circle. The particle cannot wrap around the circle to give us new states. Another example of the double strike mechanism is provided by the quantization of closed strings on an orbifold. The quantization of closed strings on the $\mathbb{R}^1/\mathbb{Z}_2$ orbifold was studied in Chapter 13. There we lost the states that are not invariant under $X \to -X$, and we gained a sector of twisted states.

17.5 Constraint and mass formula

In our earlier study of light-cone closed strings, the constraint $\alpha_0^- = \bar{\alpha}_0^-$ led to the requirement that $L_0^\perp - \bar{L}_0^\perp$ annihilates the states of the theory. This is still true in the current situation, because x^- is not compactified and, as a result, its left-moving and right-moving momenta remain equal. There is one difference, though: this time the fact that $L_0^\perp - \bar{L}_0^\perp$ vanishes does not imply that $N^\perp - \bar{N}^\perp$ vanishes. To see this, we need explicit expressions for $L_0^\perp$ and $\bar{L}_0^\perp$. These are readily obtained after considering equations (13.37) and (13.42). The modifications are minor. The sums over I split into a sum over i and another term which corresponds to the compact dimension. We find

$$\begin{aligned} \bar{L}_0^\perp &= \frac{1}{2}\bar{\alpha}_0^I\bar{\alpha}_0^I + \bar{N}^\perp = \frac{\alpha'}{4}p^i p^i + \frac{1}{2}\bar{\alpha}_0\bar{\alpha}_0 + \bar{N}^\perp, \\ L_0^\perp &= \frac{1}{2}\alpha_0^I\alpha_0^I + N^\perp = \frac{\alpha'}{4}p^i p^i + \frac{1}{2}\alpha_0\alpha_0 + N^\perp. \end{aligned} \tag{17.47}$$

The operators $N^\perp$ and $\bar{N}^\perp$ include contributions from all the oscillators: those with superscript i and those corresponding to the string coordinate X. We can now calculate

$$\begin{aligned} L_0^\perp - \bar{L}_0^\perp &= \frac{1}{2}(\alpha_0\alpha_0 - \bar{\alpha}_0\bar{\alpha}_0) + N^\perp - \bar{N}^\perp \\ &= -\alpha' pw + N^\perp - \bar{N}^\perp, \end{aligned} \tag{17.48}$$

where we made use of (17.3). It follows that the constraint on physical states takes the form

$$N^\perp - \bar{N}^\perp = \alpha' \, pw \, . \tag{17.49}$$

On states that have either zero momentum or zero winding, the number of left-moving and the number of right-moving excitations must agree. But on states that have both nonzero momentum and nonzero winding $N^\perp - \bar{N}^\perp$ cannot vanish. It must equal $\alpha' pw$. Since both $N^\perp$ and $\bar{N}^\perp$ are operators with integer eigenvalues, the left-hand side of (17.49) always takes integer values. Because of the quantization conditions (17.42) and (17.46),

$$\alpha' pw = \alpha' \cdot \frac{n}{R} \cdot \frac{mR}{\alpha'} = nm \, , \tag{17.50}$$

so the right-hand side of (17.49) also takes integer values. In terms of these integers, (17.49) takes the simple form

$$N^\perp - \bar{N}^\perp = nm \, . \tag{17.51}$$

An important tool that we have used to study the string spectrum is the formula for the mass-squared of states. To obtain such a formula we take the viewpoint of an observer that lives in the 25-dimensional Minkowski spacetime that does not include the compact dimension. For this observer $M^2 = -p^2$, where p is the 25-dimensional momentum of the states. This momentum vector has components p^+, p^-, and p^i; it does not include the component of momentum that is along the compact dimension. We therefore have

$$M^2 = 2p^+p^- - p^i p^i = \frac{2}{\alpha'}(L_0^\perp + \bar{L}_0^\perp - 2) - p^i p^i \, , \tag{17.52}$$

where we used equation (13.46) to write p^- in terms of Virasoro operators. Replacing the values of $L_0^\perp$ and $\bar{L}_0^\perp$ from (17.5) we find

$$M^2 = \frac{1}{\alpha'}(\alpha_0\alpha_0 + \bar{\alpha}_0\bar{\alpha}_0) + \frac{2}{\alpha'}(N^\perp + \bar{N}^\perp - 2) \, . \tag{17.53}$$

Our last step makes use of (17.3) to give

$$M^2 = p^2 + w^2 + \frac{2}{\alpha'}(N^\perp + \bar{N}^\perp - 2) \, . \tag{17.54}$$

This is our result for the mass-squared operator. Here p and w are the quantized momentum and winding associated with the compact dimension. The last term is familiar from the case of closed strings in Minkowski space.

Consider the contribution of the momentum p to M^2, setting for the moment all other terms to zero. This momentum then gives a rest mass, or rest energy, of $M = |p|$. The internal momentum therefore contributes to the rest energy of the string in the same way as the

momentum of a massless particle contributes to its energy. Consider now the contribution of the winding w to M^2, setting again all other terms to zero. In this case $M = |w|$, and we can provide a simple interpretation. Consider a string state that is wound $|m|$ times around the compact dimension. The length of such a string is $|m|2\pi R$, and since the string tension is $1/2\pi\alpha'$, the rest energy of the string is equal to

$$M = \frac{1}{2\pi\alpha'}\,|m|2\pi R = \frac{|m|R}{\alpha'} = |w|\,. \tag{17.55}$$

Thus the contribution of the winding to the mass is naturally understood as the energy associated with the stretching that is required to wrap the string around the compactified dimension.

When a string has both nonzero momentum p and winding w, their contributions to the rest energy of the state do not simply add. They must be added in quadrature. The number operators contribute linearly to the square of the rest energy. In some sense a state with momentum, winding, and oscillator excitations can be viewed as a bound state that is composed of various building blocks which interact to give a total rest energy that is smaller than the sum of the energies of the separate constituents.

Quick calculation 17.2 Show that the Hamiltonian of the closed string theory under study is given by

$$H = \frac{\alpha'}{2}(p^i\,p^i + p^2 + w^2) + N^\perp + \bar{N}^\perp - 2\,. \tag{17.56}$$

17.6 State space of compactified closed strings

We now construct explicitly the state space of the quantum closed string we have been studying. Let us begin with the states that have no oscillator excitations. These states have the familiar momentum labels that are associated with the 25-dimensional Minkowski space, but they also carry additional labels that specify the momentum and the winding along the compact dimension. Since the momentum is quantized as $p = n/R$, we can use the integer n as an alternative label for the momentum of the state. Similarly, with $w = mR/\alpha'$, we can use the integer m as a label for the winding of the state. Letting $\vec{p}_T$ denote the vector with components p^i, for $i = 2, \ldots, 24$, we have

$$\text{ground states:}\quad |\,p^+, \vec{p}_T;\, n, m\rangle\,, \qquad n, m \in \mathbb{Z}\,. \tag{17.57}$$

Although we refer to these as "ground states," this does not actually mean that they are all allowed states in our theory. Indeed, since $N^\perp = \bar{N}^\perp = 0$ for all of these states, the constraint (17.51) tells us that allowed ground states must have either n or m, or both, equal to zero. We call these "ground states" only to emphasize the fact that each of them is killed by all of the annihilation operators of the theory. Note also that although those states in (17.57) that have both n and m nonzero are not allowed, by themselves, to be states in

our theory, they can be acted upon by appropriate combinations of creation operators in order to produce allowed states.

A basis for the state space is constructed by applying creation operators to the above states. The general *candidate* basis state of the theory is

$$\Big[\prod_{r=1}^{\infty}\prod_{i=2}^{24}\big(a_r^{i\,\dagger}\big)^{\lambda_{i,r}}\Big]\Big[\prod_{s=1}^{\infty}\prod_{j=2}^{24}\big(\bar{a}_s^{j\,\dagger}\big)^{\bar{\lambda}_{j,s}}\Big]\Big[\prod_{k=1}^{\infty}\big(a_k{}^{\dagger}\big)^{\lambda_k}\Big]\Big[\prod_{l=1}^{\infty}\big(\bar{a}_l^{\dagger}\big)^{\bar{\lambda}_l}\Big]|\,p^+,\,\vec{p}_T;\;n,m\rangle\,. \tag{17.58}$$

We separated out the oscillators that arise from the compact dimension because they carry no 25-dimensional Lorentz index. The number operators $N^\perp$ and $\bar{N}^\perp$ act on the above state to give

$$N^\perp=\sum_{r=1}^{\infty}\sum_{i=2}^{24}r\lambda_{i,r}+\sum_{k=1}^{\infty}k\lambda_k\,,\quad \bar{N}^\perp=\sum_{s=1}^{\infty}\sum_{j=2}^{24}s\bar{\lambda}_{j,s}+\sum_{l=1}^{\infty}l\bar{\lambda}_l\,. \tag{17.59}$$

The candidate state (17.58) is a member of the state space if and only if it satisfies the constraint (17.51):

$$N^\perp-\bar{N}^\perp=nm\,. \tag{17.60}$$

The mass-squared of the state is given by (17.54):

$$M^2=\Big(\frac{n}{R}\Big)^2+\Big(\frac{mR}{\alpha'}\Big)^2+\frac{2}{\alpha'}(N^\perp+\bar{N}^\perp-2)\,. \tag{17.61}$$

In order to familiarize ourselves with the spectrum we now examine some of the closed string states in more detail.

States with $m=n=0$ These states have neither momentum nor winding in the compact dimension. One such string might be string (a) in Figure 17.1 provided, of course, that it has no net momentum along the x axis. Equation (17.60) tells us that $N^\perp=\bar{N}^\perp$, so we must be sure to match the number of left-moving and right-moving oscillators that act on the state. The vacuum state is simply

$$|p^+,\,\vec{p}_T;0,0\rangle\,,\quad M^2=-\frac{4}{\alpha'}\,. \tag{17.62}$$

This is the closed string tachyon state. Next come the massless states that arise when $N^\perp=\bar{N}^\perp=1$. Since in both the left and right sectors we have two kinds of oscillators (those that belong to the compact direction and those that do not) there are four ways we can combine the oscillators to form massless states:

$$\begin{aligned}
&a_1^\dagger\,\bar{a}_1^\dagger\;|p^+,\,\vec{p}_T;0,0\rangle\,,\\
&a_1^\dagger\bar{a}_1^{i\,\dagger}\,|p^+,\,\vec{p}_T;0,0\rangle\,,\\
&a_1^{i\dagger}\,\bar{a}_1^\dagger\,|p^+,\,\vec{p}_T;0,0\rangle\,,\\
&a_1^{i\,\dagger}\bar{a}_1^{j\dagger}|p^+,\,\vec{p}_T;0,0\rangle\,.
\end{aligned} \tag{17.63}$$

The first line contains only one state. Since it carries no 25-dimensional index, this is a state of a massless scalar field. The states in the second and third line each carry the index i, which is a complete light-cone index for the 25-dimensional spacetime. As a result, these are photon states, and each set of states corresponds to a Maxwell field. So we get a total of two Maxwell fields. This is interesting, since ordinary closed strings in Minkowski space did not even give rise to a single Maxwell field, let alone two! Finally, the states in the fourth line have exactly the same structure as the massless closed string states of Minkowski space, except that the dimensionality is now reduced to 25. Therefore these states comprise a gravity field, a Kalb–Ramond field, and a dilaton, all of them living in 25 spacetime dimensions.

If you look at the above states, you can see that all that has really happened is that the compactification reorganized the various massless states of the original 26-dimensional theory. The states now organize themselves into 25-dimensional Lorentz tensors. The number of oscillators has not changed, but since we now have one index that is inert under 25-dimensional Lorentz transformations, the types of fields have changed.

This reorganization of states upon compactification has been known in particle physics since the early work of Kaluza and Klein, who in the early 1920s attempted to build a four-dimensional unified theory of gravitation and electromagnetism by compactification of a purely gravitational theory in five dimensions. It is possible to understand heuristically what happens in this case. Let $g_{\mu\nu}$, where μ and ν are five-dimensional indices, represent the gravity field in five dimensions. Suppose that the dimension we are going to compactify is the fifth dimension, and let m and n be four-dimensional indices that run along the extended spacetime. As a matrix, $g_{\mu\nu}$ can be decomposed into a matrix g_{mn}, which represents the four-dimensional gravity field, a vector g_{m5}, which represents a four-dimensional Maxwell field (g_{5m} is not a new field since the original metric is symmetric), and a single component g_{55}, which represents a four-dimensional scalar. This is the main result of Kaluza and Klein: a five-dimensional gravity theory, after compactifying down to four dimensions, produces a gravity theory that is coupled to a Maxwell field and to a massless scalar.

How is it that Kaluza and Klein were only able to produce one Maxwell field from their compactified theory of gravity, while our list in (17.6) gave two? This is possible because string theory is more than simply a theory of gravity. The second Maxwell field arises from the higher-dimensional Kalb–Ramond field. So, even in string theory, only one Maxwell field arises from the gravitational field. It is just that now the compactification affects other fields in addition to gravity. If $B_{\mu\nu}$ denotes a five-dimensional Kalb–Ramond field, then, upon compactification, B_{m5} represents a four-dimensional Maxwell field. Indeed, forming linear combinations of the states in the second and third lines of (17.6) we have

$$\begin{aligned}&\left(a_1^\dagger \bar{a}_1^{i\dagger} + a_1^{i\dagger} \bar{a}_1^\dagger\right)|p^+, \vec{p}_T; 0, 0\rangle\,,\\ &\left(a_1^\dagger \bar{a}_1^{i\dagger} - a_1^{i\dagger} \bar{a}_1^\dagger\right)|p^+, \vec{p}_T; 0, 0\rangle\,.\end{aligned} \tag{17.64}$$

The states in the first line correspond to photon states that arise from the 26-dimensional graviton states. The states in the second line correspond to photon states that arise from the 26-dimensional Kalb–Ramond states.

States with $n = 0$ or with $m = 0$ Since these states have either nonzero momentum or winding, but not both, they must still satisfy $N^\perp = \bar{N}^\perp$. The ground states are

$$\begin{aligned} &|p^+, \vec{p}_T; n, 0\rangle\,, \quad M^2 = \frac{n^2}{R^2} - \frac{4}{\alpha'}\,, \\ &|p^+, \vec{p}_T; 0, m\rangle\,, \quad M^2 = \frac{m^2 R^2}{\alpha'^2} - \frac{4}{\alpha'}\,. \end{aligned} \tag{17.65}$$

Because they have no 25-dimensional Lorentz indices, both sets of states correspond to scalar fields. For any fixed n or m, the states may be tachyonic, massless, or massive, depending on the value of the radius R. Acting with oscillators on these vacua produces heavier states. Such states have $N^\perp + \bar{N}^\perp \geq 2$ and, as a result, they are massive for all values of the radius R. It is therefore impossible to find massless vectors in this sector of the state space.

States with $n = m = \pm 1$ or $n = -m = \pm 1$ These states have both momentum and winding, so $N^\perp$ and $\bar{N}^\perp$ must be different. The two situations we consider are

$$\begin{aligned} n = \quad m = \pm 1 &\longrightarrow N^\perp - \bar{N}^\perp = \quad 1\,, \\ n = -m = \pm 1 &\longrightarrow N^\perp - \bar{N}^\perp = -1\,. \end{aligned} \tag{17.66}$$

The lowest-mass solutions of $N^\perp - \bar{N}^\perp = 1$ arise when $N^\perp = 1$ and $\bar{N}^\perp = 0$. There are two kinds of states satisfying this condition:

$$\begin{aligned} &a_1^\dagger\, |p^+, \vec{p}_T, \pm 1, \pm 1\rangle\,, \\ &a_1^{i\dagger}\, |p^+, \vec{p}_T, \pm 1, \pm 1\rangle\,. \end{aligned} \tag{17.67}$$

Similarly, the lowest-mass solutions of $N^\perp - \bar{N}^\perp = -1$ arise when $N^\perp = 0$ and $\bar{N}^\perp = 1$, so these are

$$\begin{aligned} &\bar{a}_1^\dagger\, |p^+, \vec{p}_T; \pm 1, \mp 1\rangle\,, \\ &\bar{a}_1^{i\dagger}\, |p^+, \vec{p}_T; \pm 1, \mp 1\rangle\,. \end{aligned} \tag{17.68}$$

Both groups of states above have mass

$$M^2(R) = \frac{1}{R^2} + \frac{R^2}{\alpha'^2} - \frac{2}{\alpha'} = \left(\frac{1}{R} - \frac{R}{\alpha'}\right)^2. \tag{17.69}$$

It is interesting to note that there is a particular radius R^* at which $M^2(R^*) = 0$, so that all of the states are massless:

$$\frac{1}{R^*} = \frac{R^*}{\alpha'} \;\rightarrow\; R^* = \sqrt{\alpha'}\,. \tag{17.70}$$

R^* is precisely the string length. At this radius, which is called the self-dual radius (for reasons that will become clear later), the interpretation of the above states is simple. The states in the first rows of (17.67) and (17.68) comprise a total of four massless scalars. The states in the second rows of (17.67) and (17.68) comprise a total of four *massless gauge fields*.

Since all of these states have nonzero winding, they are truly "stringy" states that could not arise in a theory of particles.

Quick calculation 17.3 Prove that in the sector of states with $n \neq 0$ and $m \neq 0$ there are no additional states that can ever become massless.

Previously, we identified two Maxwell fields in the sector where momentum and winding are both zero (see (17.6)). These two $U(1)$ fields (using the language introduced in Section 15.3) arise from the gravity and Kalb–Ramond fields in the 26-dimensional theory via a mechanism that applies both in particle theory and in string theory. Now we have obtained four additional gauge fields, all of them massless at the self-dual radius and all of them truly "stringy." Something interesting now occurs, whose description uses the two physically equivalent $U(1)$ fields corresponding to the sum and differences of the states (17.6). Each of these latter $U(1)$ fields happens to combine with two of the stringy gauge fields to form a set of three gauge bosons that interact in a way described by an $SU(2)$ Yang–Mills theory. More explicitly, the combinations are

$$\begin{aligned} &a_1^\dagger\, \bar{a}_1^{i\dagger}\, |p^+, \vec{p}_T; 0, 0\rangle \quad \text{with} \quad \bar{a}_1^{i\dagger}\, |p^+, \vec{p}_T; \pm 1, \mp 1\rangle\,, \\ &a_1^{i\dagger}\, \bar{a}_1^\dagger\, |p^+, \vec{p}_T; 0, 0\rangle \quad \text{with} \quad a_1^{i\dagger}\, |p^+, \vec{p}_T, \pm 1, \pm 1\rangle\,. \end{aligned} \tag{17.71}$$

In the compactification of a *particle* theory of gravity and Kalb–Ramond fields the resulting theory would have a $U(1) \times U(1)$ gauge group. In string theory, the $U(1) \times U(1)$ gauge group gets enhanced to an $SU(2) \times SU(2)$ symmetry at the self-dual radius. We have previously seen how Yang–Mills theories arise on the world-volume of coincident D-branes. Now we see one way in which Yang–Mills theories can appear in closed string theory.

17.7 A striking spectrum coincidence

We have already seen that the mass spectrum of the compactified string is very much dependent on the radius of compactification. At the self-dual radius R^*, for example, we get some additional massless gauge fields. Now we will discover a surprising property of the spectrum. To bring this property out into the open, we must look into equations (17.61) and (17.60), which read

$$M^2 = \frac{n^2}{R^2} + \frac{m^2 R^2}{\alpha'^2} + \frac{2}{\alpha'}(N^\perp + \bar{N}^\perp - 2)\,, \qquad N^\perp - \bar{N}^\perp = nm\,. \tag{17.72}$$

Now here is the remarkable property: the closed string spectrum for a compactification with radius R is identical to the closed string spectrum for a compactification with radius $\tilde{R} = \alpha'/R$. We will verify this property shortly, but it turns out that even more is true: the two compactifications are physically indistinguishable. This is the T-duality of closed string theory. Note that this means that, in closed string theory, a compactification with extremely large radius is equivalent to a compactification with extremely small radius! The radii R and α'/R are called dual radii:

$$R \longleftrightarrow \frac{\alpha'}{R} \equiv \tilde{R}\,. \tag{17.73}$$

To verify the coincidence of the spectrum, let us write an expression for the mass-squared for each of these radii:

$$\begin{aligned} M^2\Big(R;\ n,m\Big) &= \frac{n^2}{R^2} + \frac{m^2R^2}{\alpha'^2} + \frac{2}{\alpha'}(N^\perp + \bar{N}^\perp - 2)\,, \\ M^2\Big(\tilde{R};\ n,m\Big) &= \frac{n^2R^2}{\alpha'^2} + \frac{m^2}{R^2} + \frac{2}{\alpha'}(N^\perp + \bar{N}^\perp - 2)\,. \end{aligned} \tag{17.74}$$

These do not look the same, but the difference is merely superficial. As n and m run over all possible integers, the *lists* of masses that result are the same. More explicitly, for all $n, m \in \mathbb{Z}$,

$$M^2\Big(R;\ n,m\Big) = M^2\Big(\tilde{R};\ m,n\Big)\,. \tag{17.75}$$

In writing this equality, of course, we are comparing states with identical oscillator structure, otherwise the contributions from the number operators would not agree. Note also that the exchange of n and m does not affect the constraint in (17.72). This proves that the mass spectra of theories with dual radii are identical. The key to this equality is the opposite dependences on the radius of the contributions to the mass-squared from the momentum and the winding. The exchange of n and m in (17.75) is just the exchange of winding and momentum quantum numbers.

In the following section we will give evidence that dual radii are actually physically indistinguishable. This is a property of string theory that has been proven beyond doubt. The special radius R^* in (17.70) is the unique radius that is mapped to itself under the transformation in (17.73). The duality then implies that each radius smaller than R^* is equivalent to some radius larger than R^*. In this sense, the radius R^* represents the minimal radius that can be attained in toroidal compactification.

In the present analysis the value of the radius of the circle is adjustable. The theory did not select for us a particular radius the way it selected, for example, a particular spacetime dimension. The radius of the circle must be viewed as a parameter of a particular class of compactified spacetimes that allows a consistent definition of string theory. The compactification radius is *not* a parameter of string theory itself, but rather a parameter of a spacetime allowed in string theory. It is an adjustable parameter. Such parameters are sometimes called *moduli*, and the set of values the parameters can take is called the *moduli space*. What we learn from T-duality is that the moduli space of compactifications into a circle can be taken to include only radii larger than or equal to R^*.

To appreciate how a very small circle can appear to be very large, consider the following heuristic argument. On a circle of radius R momentum is quantized in units of $1/R$. If R is very large the spacing between momentum eigenvalues is very small and the spectrum is almost continuous. When the radius is very small, the momentum eigenvalues are largely spaced. In standard particle theory, this is a signal of a small circle. In string theory the situation is different. The winding eigenvalues are quantized in units of R/α', and as

the radius R becomes very small, the winding spectrum becomes almost continuous. An observer seeing this continuum might conclude that the compact dimension is very large.

17.8 Duality as a full quantum symmetry

We have seen that T-duality is a symmetry of the mass spectrum of compactified closed string theory. This result, by itself, does not imply that the physics is identical at dual radii. In this section we will prove that the full theory of *free* closed strings is T-duality invariant. It can be proven that T-duality holds even when interactions between strings are included, but we will not attempt to do so in this book.

How do we go about showing the equivalence of the dual theories? We will first bring out into the open some additional structure that exists in compactified string theory. We then explain the equivalence of dual theories in two related ways. In the first we show that T-duality arises as an interpretative ambiguity of a *single* theory: one possible choice of string coordinate suggests that the radius of the circle is R, while another, physically equivalent choice, suggests that the radius of the circle is α'/R. In the second, T-duality is exhibited as an equivalence between two distinct theories. We consider a theory at radius R and a theory at radius α'/R. We then find a one-to-one correspondence between the operators of these theories that preserves all the commutation relations and takes one Hamiltonian into the other. This is the same strategy that we used to demonstrate the duality symmetry of the harmonic oscillator back in Section 17.1.

The additional structure of compactified string theory is related to the coordinate zero mode q_0 that appeared both in X_L and in X_R but did not appear in the full coordinate $X = X_L + X_R$ (see (17.3)). Recall that the winding operator w had a vanishing commutator with all the operators that appear in X. It is natural to make (q_0, w) into a conjugate pair of variables. To this end, we now introduce a dual "coordinate" operator, defined by

$$\tilde{X}(\tau,\sigma) \equiv X_L(\tau+\sigma) - X_R(\tau-\sigma)\,. \tag{17.76}$$

Making use of (17.3), we find that this coordinate takes the form

$$\tilde{X}(\tau,\sigma) = q_0 + \alpha' w\,\tau + \alpha' p\,\sigma + i\sqrt{\frac{\alpha'}{2}}\sum_{n\neq 0}\frac{e^{-in\tau}}{n}(\bar{\alpha}_n e^{-in\sigma} - \alpha_n e^{in\sigma}). \tag{17.77}$$

This coordinate is interesting because in it p multiplies σ and w multiplies τ, which reverses the familiar pairings in X. Moreover, q_0 appears in place of x_0. It follows that the momentum associated with the coordinate q_0 is w (since it appears with τ), and the winding associated with q_0 is p (since it appears with σ). The coordinate $\tilde{X}$ brings q_0 into play. The precise meaning of q_0 will be explained below.

We can now define a momentum $\tilde{\mathcal{P}}^\tau$ conjugate to the coordinate $\tilde{X}$:

$$\tilde{\mathcal{P}}^\tau \equiv \frac{1}{2\pi\alpha'}\,\partial_\tau \tilde{X} = \frac{1}{2\pi\alpha'}(\dot{X}_L - \dot{X}_R)\,. \tag{17.78}$$

We postulate the commutator

$$[\,\tilde{X}(\tau,\sigma)\,,\tilde{\mathcal{P}}^{\tau}(\tau,\sigma')\,]=i\delta(\sigma-\sigma') \tag{17.79}$$

and demand that the commutator between two coordinates or between two momenta vanishes. You may be concerned because these commutators involve the same oscillators that we encountered before, so that their commutation relations are already determined. This is true, but the same commutation relations emerge, because the only difference between the pairs $(X,\mathcal{P}^{\tau})$ and $(\tilde{X},\tilde{\mathcal{P}}^{\tau})$ is that *the sign of X_R is reversed*. As you can see from (17.3), this sign reversal is implemented by changing the sign of all the α_n oscillators, exchanging x_0 and q_0, and exchanging p and q. These changes do not affect the commutators (17.35), which are therefore reproduced. The pair (q_0,w) appears in $\tilde{X}$ just as the pair (x_0,p) appears in X, so we find

$$[\,q_0,w]=i\,, \tag{17.80}$$

as well as

$$[q_0,p]=[q_0,\alpha_n]=[q_0,\bar{\alpha}_n]=0\,,\quad n\neq 0\,. \tag{17.81}$$

We know that x_0 is a coordinate that lives on a circle of radius R, because the associated canonical momentum p is quantized with eigenvalues n/R. Similarly, from the quantization $w=mR/\alpha'$ of the winding momentum, we infer that the associated coordinate q_0 lives on a circle of radius $\tilde{R}=\alpha'/R$. Thus the string coordinate $\tilde{X}$ itself is a coordinate on a circle of radius α'/R. The relation $[q_0,w]=i$ is formal in the same way that $[x_0,p]=i$ is. Well defined operators analogous to (17.43) are given by

$$e^{i\ell q_0/\tilde{R}}\,,\quad \ell\in\mathbb{Z}. \tag{17.82}$$

We also have

$$e^{-i\ell q_0/\tilde{R}}\,w\,e^{i\ell q_0/\tilde{R}}=w+\frac{\ell}{\tilde{R}}\,. \tag{17.83}$$

The operators (17.82) have a well defined action on the spectrum.

Quick calculation 17.4 Verify that acting on a state with winding number m, the operator $e^{i\ell q_0/\tilde{R}}$ gives a state with winding number $m+\ell$.

The only commutator that is not fixed is that of x_0 with q_0. In terms of well defined operators we declare

$$\left[\,e^{-i\ell x_0/R}\,,\,e^{-imq_0/\tilde{R}}\,\right]=0\,,\quad \ell,m\in\mathbb{Z}\,. \tag{17.84}$$

This is consistent with our understanding that the two kinds of operators involved here simply act on momentum or winding labels independently of each other. Naively, we write $[x_0,q_0]=0$.

The Hamiltonian derived from $(\tilde{X}\,,\tilde{\mathcal{P}}^{\tau})$ coincides with the one derived from $(X\,,\mathcal{P}^{\tau})$. This is clear from equation (17.56): the exchange of p and w has no effect, and the minus sign in the α_n oscillators does not affect the number operators. T-duality emerges as an *interpretative ambiguity*. We began with a theory that used the coordinate X to describe a

radius R compactification. When the operator content of the theory is examined, we find a different coordinate $\tilde{X}$ which gives an equally compelling interpretation of the theory as one for which the compactification radius is α'/R. In both interpretations we get the same Hamiltonian. In summary, duality arises because of the possibility of replacing X with $\tilde{X}$:

$$\text{T-duality:} \quad X = X_L + X_R \longrightarrow \tilde{X} = X_L - X_R \,. \tag{17.85}$$

This expression for the coordinate $\tilde{X}$ dual to X will be helpful when we discuss open strings in Chapter 18.

To describe duality as a map between two theories, we attempt to formulate the passage from $(X\,,\mathcal{P}^\tau)$ operators to $(\tilde{X}\,,\tilde{\mathcal{P}}^\tau)$ operators as a map:

$$(X\,,\mathcal{P}^\tau) \longrightarrow (\tilde{X}\,,\tilde{\mathcal{P}}^\tau)\,. \tag{17.86}$$

It is clear from (17.34) and (17.79) that the commutation relations are preserved. From our earlier comments, (17.86) is equivalent to the map

$$(X_L\,,X_R) \to (X_L\,,-X_R)\,. \tag{17.87}$$

Finally, by inspection of (17.3), we see that this map is implemented by the following map of oscillators and zero modes:

$$\left\{\begin{array}{c} x_0 \to q_0 \\ q_0 \to x_0 \end{array}\right\} \quad \left\{\begin{array}{c} p \to w \\ w \to p \end{array}\right\} \quad \left\{\begin{array}{c} \alpha_n \to -\alpha_n \\ \bar{\alpha}_n \to \quad \bar{\alpha}_n \end{array}\right\}. \tag{17.88}$$

This is not a map between two theories, but it will help us to construct one.

To find such a map, we now consider two distinct theories: one for which the familiar interpretation is that the compactification radius is R and another for which the familiar interpretation is that the compactification radius is α'/R. Our task is then to produce a map between the operators of the two theories that is consistent with the commutation relations and takes one Hamiltonian into the other. One important point to be aware of is that the map must respect the operator quantization conditions.

In Table 17.1 we have recorded the operators, quantization conditions, and commutation relations for a theory with radius R. Similarly, in Table 17.2 we have recorded the operators, quantization conditions, and commutation relations for a theory with radius α'/R. In this table we have placed a tilde over all of the operators in order to distinguish them from the operators of the other theory. The requisite operator map is suggested by (17.88) but now takes the form

$$\left\{\begin{array}{c} x_0 \to \tilde{q}_0 \\ q_0 \to \tilde{x}_0 \end{array}\right\} \quad \left\{\begin{array}{c} p \to \tilde{w} \\ w \to \tilde{p} \end{array}\right\} \quad \left\{\begin{array}{c} \alpha_n \to -\tilde{\alpha}_n \\ \bar{\alpha}_n \to \quad \tilde{\bar{\alpha}}_n \end{array}\right\}. \tag{17.89}$$

For all operators associated with the 25-dimensional spacetime the map is the identity map. The transformations (17.89) respect the commutation relations and map one Hamiltonian into the other. Moreover, we see explicitly that they map operators to others that live on

Table 17.1 A theory where the X coordinate lives on a circle of radius R; listed are the Hamiltonian, the zero modes, and a list of commutation relations

Theory with radius R
$H(R) = \frac{1}{2}\alpha' p^i p^i + \frac{1}{2}\alpha'(p^2 + w^2) + N^\perp + \bar{N}^\perp - 2$
x_0 lives on a circle of radius R
p has eigenvalues n/R
q_0 lives on a circle of radius α'/R
w has eigenvalues mR/α'
$[x_0, p] = [q_0, w] = i$
$[\bar{\alpha}_m, \bar{\alpha}_n] = [\alpha_m, \alpha_n] = m\delta_{m+n,0}$
$[\alpha_m, \bar{\alpha}_n] = 0$

Table 17.2 A theory where the X coordinate lives on a circle of radius α'/R; listed are the Hamiltonian, the zero modes, and a list of commutation relations

Theory with radius $\tilde{R} = \alpha'/R$
$\tilde{H}(\tilde{R}) = \frac{1}{2}\alpha' p^i p^i + \frac{1}{2}\alpha'(\tilde{p}^2 + \tilde{w}^2) + \tilde{N}^\perp + \tilde{\bar{N}}^\perp - 2$
$\tilde{x}_0$ lives on a circle of radius α'/R
$\tilde{p}$ has eigenvalues mR/α'
$\tilde{q}_0$ lives on a circle of radius R
$\tilde{w}$ has eigenvalues n/R
$[\tilde{x}_0, \tilde{p}] = [\tilde{q}_0, \tilde{w}] = i$
$[\tilde{\bar{\alpha}}_m, \tilde{\bar{\alpha}}_n] = [\tilde{\alpha}_m, \tilde{\alpha}_n] = m\delta_{m+n,0}$
$[\tilde{\alpha}_m, \tilde{\bar{\alpha}}_n] = 0$

similar spaces and have identical spectra. Both x_0 and $\tilde{q}_0$, for example, live on identical circles. Both p and $\tilde{w}$ have the same spectrum. The map of oscillators includes a sign factor that does not affect $N^\perp$. This map establishes the physical equivalence of the theories under consideration, and it proves that T-duality is an exact symmetry of free closed string theory compactified on a circle.

Problems

Problem 17.1 Zero mode Hamiltonian.

We can use an action like (12.81) to study the dynamics of the compact coordinate X:

$$S = \frac{1}{4\pi\alpha'} \int d\tau \int_0^{2\pi} d\sigma \left(\dot{X}\,\dot{X} - X'\,X' \right).$$

Consider now the zero mode expansion of the string coordinate in the sector of winding number m:

$$X(\tau,\sigma) = x(\tau) + mR\sigma \,.$$

Find the action for $x(\tau)$. Calculate the Hamiltonian, and show that you recover the contributions to (17.56) arising from the compact dimension.

Problem 17.2 Counting massless gauge fields.

Consider a string compactification where k coordinates are made into circles of critical radius. Describe the candidate ground states of this theory, and give the expression for the Hamiltonian. Find the number of massless vector fields that arise in the lower-dimensional spacetime. Give the explicit list of states corresponding to these fields.

Problem 17.3 Charge carried by winding strings.

The zero mode description of a string with winding number ℓ is given by

$$X^m(\tau,\sigma) = x^m(\tau)\,, \quad X(\tau,\sigma) = \ell R\sigma\,, \tag{1}$$

where the index m is used to denote all string coordinates except for $X^{25} \equiv X$. Consider the coupling term (16.3) of the Kalb–Ramond field to the string:

$$S = -\int d\tau d\sigma\ \frac{\partial X^\mu}{\partial\tau}\frac{\partial X^\nu}{\partial\sigma}\,B_{\mu\nu}(X)\ .$$

Calculate the terms in S that couple $B_{m,25} = -B_{25,m}$ to the string trajectory $x^\alpha(\tau)$. The field $B_{m,25}$ plays the role of a Maxwell field in the 25-dimensional space, as explained in Section 17.6. Conclude that the winding string state described by (1) carries an electric charge proportional to ℓ.

Problem 17.4 Compactification on T^2 with a constant Kalb–Ramond field.

Assume that x^2 and x^3 are each compactified into a circle of radius R. The corresponding string coordinates are called X^r, with $r = 2, 3$. Moreover, there is a nonvanishing Kalb–Ramond field with expectation value

$$B_{23} \equiv \frac{1}{2\pi\alpha'}\,b\,, \tag{1}$$

where b is a dimensionless constant. All other components of $B_{\mu\nu}$ vanish.

(a) Build an action for the $X^r(\tau,\sigma)$ by adding an action of the type used in Problem 17.1 to the action in Problem 17.3.

(b) Consider the following expansion for the zero mode part of the coordinates:

$$X^r = x^r(\tau) + m_r R\,\sigma\,, \quad r = 2, 3. \tag{2}$$

Show that the Lagrangian for $x^r(\tau)$ is

$$L = \frac{1}{2\alpha'}\left((\dot{x}^2)^2 + (\dot{x}^3)^2\right) - \frac{\alpha'}{2}\left((w_2)^2 + (w_3)^2\right) - b\,\left(\dot{x}^2 w_3 - \dot{x}^3 w_2\right)\,. \tag{3}$$

Here $w_r = m_r R/\alpha'$, as usual. The last term in (3) is a total derivative, but it is important in the quantum theory, as you will see.

(c) Define momenta canonical to x^r, compute the Hamiltonian, and show that it takes the form

$$H = \frac{\alpha'}{2}\Big((p_2 + bw_3)^2 + (p_3 - bw_2)^2 + w_2^2 + w_3^2\Big)\,. \tag{4}$$

Verify that the Hamiltonian generates the correct equations of motion. Note that the quantization conditions on the momenta are $p_r = n_r/R$.

(d) While we have only looked explicitly at the zero modes, the oscillator expansion of the coordinates works just as before. Write the appropriate expansions for the coordinates $X^2(\tau,\sigma)$ and $X^3(\tau,\sigma)$ along the lines of equation (17.32). Explain why the mass-squared operator takes the form

$$M^2 = \Big(\frac{n_2}{R} + b\,\frac{m_3 R}{\alpha'}\Big)^2 + \Big(\frac{n_3}{R} - b\,\frac{m_2 R}{\alpha'}\Big)^2 + \Big(\frac{m_2 R}{\alpha'}\Big)^2 + \Big(\frac{m_3 R}{\alpha'}\Big)^2 + \frac{2}{\alpha'}\Big(N^\perp + \bar N^\perp - 2\Big)\,. \tag{5}$$

(e) Show that the constraint $L_0^\perp - \bar L_0^\perp = 0$ yields

$$N^\perp - \bar N^\perp = n_2 m_2 + n_3 m_3\,. \tag{6}$$

Problem 17.5 Dualities in the T^2 compactification with Kalb–Ramond field.

In Problem 17.4 you obtained equations (5) and (6), which define the spectrum of a compactification on a square torus T^2 with radius R and with Kalb–Ramond field $B_{23} = b/(2\pi\alpha')$. Show that the spectrum is unchanged under the following duality transformations on the background parameters R and b.

(a) The value of b is changed as

$$b \;\to\; b' = b + \ell\,\frac{\alpha'}{R^2}\,, \quad \ell \in \mathbb{Z}\,, \tag{1}$$

while R is left unchanged. Equation (1) states that b is a *periodic* variable. Let A denote the area of the torus. Use (1) to show that the "flux" parameter $f_B = B_{23}A$ is an *angle* variable (i.e., $f_B \sim f_B + 2\pi$). To prove that the spectrum is unchanged, you must find an appropriate compensating change in the quantum numbers, as in (17.75), for example. [Hint: $n_2 \to n_2 - \ell m_3$ is one needed change.]

(b) The values of R and b are changed as

$$R \;\to\; R' = \frac{\alpha'}{R}\,\frac{1}{\sqrt{1+b^2}}\,, \quad b \;\to\; b' = -b\,. \tag{2}$$

When $b = 0$ this is the familiar T-duality transformation of the radius. The compensating change here is the expected $n_r \leftrightarrow m_r$. [Hint: the algebra is easier if you first expand (5).]

(c) $b \to -b$, with R unchanged. Use $m_r \to -m_r$ as the compensating change of quantum numbers. What additional change must be done regarding oscillators to make (6) work?

The identifications in (1) and (2) are given a geometrical interpretation in Problem 26.3.

18 T-duality of open strings

T-duality relates a world in which a spatial coordinate on a Dp-brane is stretched around a circle to a different looking, but equivalent, world in which a D$(p-1)$-brane has a fixed position on a circle of dual radius. In the first world open strings can have momentum along the circle but cannot wind around it, while in the second world they have no momentum along the dual circle but, as we will see, they can in fact wind around it. We use Maxwell gauge transformations to show that, on a circle, the values of the gauge field line integral $\oint A dx$ are periodically identified. The holonomy of the gauge field along a Dp-brane direction that is wrapped on a circle is related by T-duality to the angular position of a D$(p-1)$-brane on the dual circle.

18.1 T-duality and D-branes

Let us consider the propagation of open strings in a spacetime in which one spatial dimension has been curled up into a circle. Assume that we have a space-filling D25-brane, so that the open string endpoints are free to move all over space. As before, we choose the x^{25} dimension to be compactified:

$$x^{25} \sim x^{25} + 2\pi R. \tag{18.1}$$

All open string coordinates, including X^{25}, satisfy Neumann boundary conditions at both endpoints, so they are all of NN type. In the presence of a compact dimension, closed strings exhibit fundamentally new states: closed strings can wrap around the compact dimension so that they cannot be shrunk to a point. Open strings on a space-filling D-brane have endpoints that are free to move anywhere, so open strings can always be shrunk away. Open strings exhibit no fundamentally new states in the presence of a compact dimension. The open string momentum in the x^{25} direction is quantized: $p^{25} = n/R$. This contributes an amount n^2/R^2 to the mass-squared of the string. There is no winding number w^{25}.

Now consider an open string in a related spacetime, which also has a space-filling D25-brane and a compactified x^{25}, but in which the radius of compactification is $\tilde{R} = \alpha'/R$. T-duality of closed strings tells us that the physics of these two spacetimes is indistinguishable as far as closed strings are concerned. In this new spacetime, however, open strings have their momentum quantized as $p^{25} = nR/\alpha'$, and therefore the mass-squared gets a contribution equal to n^2R^2/α'^2. It is clear that the spectrum of this open string theory does

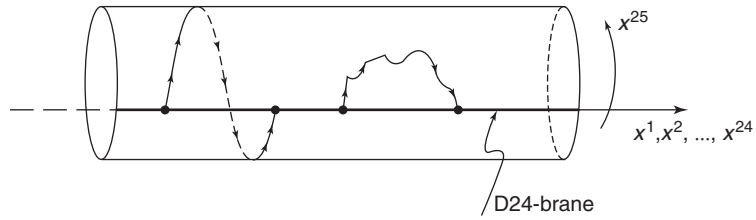

Fig. 18.1 A D24-brane, represented schematically, with x^{25} a normal direction that is curled up into a circle. The open string shown to the left wraps around the compact direction and is non-contractible since its endpoints must remain fixed on the D24-brane. The open string to the right can be contracted.

not coincide with that of the original open string theory on a circle of radius R. This implies that open strings on a D25-brane can tell the difference between a compactification with radius R and another with radius $\tilde{R} = \alpha'/R$. The inclusion of open strings seems to throw a monkey-wrench into the idea of T-duality.

But there is a solution that preserves T-duality, even in the presence of open strings. We will see that T-duality relates the spacetime with compactification radius R and a D25-brane to a spacetime with compactification radius $\tilde{R} = \alpha'/R$ and a D24-brane! The physics is equivalent for both open and closed strings if the D25-brane in the original world is replaced by a D24-brane in the dual world. In the dual world, x^{25} is the Dirichlet direction for the D24-brane, and the corresponding open string coordinate is of DD type. By convention, we set $x^{25} = 0$ to be the position of the brane along the compact direction.

In the dual world all open string endpoints must remain attached to points with $x^{25} = 0$. As a result, there are new open string configurations that cannot be contracted away. An open string that stretches from $x^{25} = 0$ up to $x^{25} = 2\pi\tilde{R}$, for example, winds around the compact direction once. This string cannot be contracted because the endpoints are not free to move along the compact dimension (see Figure 18.1). Open strings can wind any number of times, just as closed strings do. Winding open strings resemble closed strings, but they are not closed: the open string endpoints need not coincide. They typically lie on different points on the D24-brane.

When we had a D25-brane and an x^{25} circle of radius R, the open string had quantized momentum p^{25} but no winding. After the duality transformation we have a D24-brane and an x^{25} circle of radius $\tilde{R}$. The Dirichlet boundary condition imposes a zero momentum constraint, but the open string now has winding. The open string spectrum of the two theories coincide when $\tilde{R} = \alpha'/R$, because the momentum states contribute to M^2 in the first theory in the same way as the open string winding states contribute to M^2 in the second theory. By allowing the duality transformation to modify the D-brane, we can preserve T-duality in the presence of open strings. In summary,

$$\text{T-duality along } x^{25}\text{:} \quad \big(\text{D25}\,;\ R\big) \longrightarrow \big(\text{D24}\,;\ \tilde{R} = \alpha'/R\big). \tag{18.2}$$

Since the closed string spectrum is not affected by the D-branes, we have a complete physical equivalence.

To show how this works explicitly, we begin by recalling the expansion (12.32) of an NN-type open string coordinate. For $X^{25}(\tau,\sigma) \equiv X(\tau,\sigma)$, we write

$$X(\tau,\sigma) = x_0 + \sqrt{2\alpha'}\,\alpha_0\tau + i\sqrt{2\alpha'}\sum_{n\neq 0}\frac{1}{n}\alpha_n \cos n\sigma\, e^{-in\tau}. \tag{18.3}$$

We also have

$$\alpha_0 = \sqrt{2\alpha'}\,p = \sqrt{2\alpha'}\,\frac{n}{R}, \tag{18.4}$$

since the momentum on the circle is quantized. The Hamiltonian for this open string is

$$H = L_0^{\perp} - 1 = \frac{1}{2}\alpha_0^I\alpha_0^I + N^{\perp} - 1 = \alpha' p^i\, p^i + \frac{1}{2}\alpha_0\alpha_0 + N^{\perp} - 1, \tag{18.5}$$

where $i = 2, \ldots, 24$, and $N^{\perp}$ includes the contribution from the oscillators α_n^i and α_n. We now separate the string coordinate X into left-moving and right-moving components:

$$X(\tau,\sigma) = X_L(\tau+\sigma) + X_R(\tau-\sigma), \tag{18.6}$$

where

$$\begin{aligned} X_L &= \frac{1}{2}(x_0+q_0) + \sqrt{\frac{\alpha'}{2}}\alpha_0(\tau+\sigma) + \frac{i}{2}\sqrt{2\alpha'}\sum_{n\neq 0}\frac{1}{n}\alpha_n e^{-in\tau}e^{-in\sigma}, \\ X_R &= \frac{1}{2}(x_0-q_0) + \sqrt{\frac{\alpha'}{2}}\alpha_0(\tau-\sigma) + \frac{i}{2}\sqrt{2\alpha'}\sum_{n\neq 0}\frac{1}{n}\alpha_n e^{-in\tau}e^{+in\sigma}. \end{aligned} \tag{18.7}$$

The constant q_0 is arbitrary. Inspired by closed string T-duality, where we reversed the sign of the right movers in (17.85), we now define

$$\tilde{X}(\tau,\sigma) \equiv X_L - X_R, \tag{18.8}$$

and then find

$$\tilde{X}(\tau,\sigma) = q_0 + \sqrt{2\alpha'}\,\alpha_0\,\sigma + \sqrt{2\alpha'}\sum_{n\neq 0}\frac{1}{n}\alpha_n e^{-in\tau}\sin n\sigma. \tag{18.9}$$

This is, in fact, the expansion for a string that stretches from one D-brane to another. We recall equations (15.45) and (15.47), which give:

$$X^a(\tau,\sigma) = \bar{x}_1^a + \sqrt{2\alpha'}\,\alpha_0^a\,\sigma + \sqrt{2\alpha'}\sum_{n\neq 0}\frac{1}{n}\alpha_n^a\, e^{-in\tau}\sin n\sigma, \tag{18.10}$$

together with

$$\sqrt{2\alpha'}\alpha_0^a = \frac{1}{\pi}(\bar{x}_2^a - \bar{x}_1^a). \tag{18.11}$$

The coordinate difference $\bar{x}_2^a - \bar{x}_1^a$ is the separation between the D-branes in the a-direction. Equations (18.9) and (18.10) are in full correspondence. If we delete

the superscript a from (18.10) and identify the constants $\bar{x}_1$ and q_0, we recover the expansion in (18.9).

Before giving the physical interpretation of the new coordinate $\tilde{X}$, we explain why the duality $X \to \tilde{X}$ is a symmetry of the open string theory. Our work with X^a and $\mathcal{P}^a$ in Section 15.3 proves that $\tilde{X}$ and $\tilde{\mathcal{P}} = \partial_\tau \tilde{X}/(2\pi\alpha')$ satisfy the canonical commutation relations. Therefore the duality transformation does not alter the commutation relations. In addition, the Hamiltonian is unchanged. The Hamiltonian for the sector including X^a is obtained from (15.49):

$$H = 2\alpha' p^+ p^- = \alpha' p^i p^i + \frac{1}{2}\alpha_0^a \alpha_0^a + \sum_{n=1}^{\infty}\left[\alpha_{-n}^i \alpha_n^i + \alpha_{-n}^a \alpha_n^a\right] - 1. \tag{18.12}$$

If we delete the superscript a, the correspondence implies that this is the Hamiltonian that arises from $\tilde{X}$ and $\tilde{\mathcal{P}}$ together with the other string coordinates. It is then clear that the Hamiltonian coincides with the one given earlier in (18.5).

We can now turn to the physical interpretation. The string coordinate $\tilde{X}$ is of DD type, since the endpoints are fixed: $\partial_\tau \tilde{X} = 0$ for $\sigma = 0$ and $\sigma = \pi$. When σ goes from 0 to π, the open string stretches an interval

$$\tilde{X}(\tau,\pi) - \tilde{X}(\tau,0) = \sqrt{2\alpha'}\,\alpha_0\,(\pi - 0) = 2\pi\alpha' p = 2\pi\frac{\alpha'}{R}\,n = 2\pi\tilde{R}\,n. \tag{18.13}$$

Since n can take all possible integer values, the picture that emerges is that of an infinite collection of D24-branes with a uniform spacing $2\pi\tilde{R}$ along the x^{25} direction. Such a configuration is indeed physically equivalent to a single D24-brane at some fixed position on a circle of radius $\tilde{R}$.

It is interesting to note that duality interchanges boundary conditions. We have

$$\partial_\sigma X = X_L'(\tau+\sigma) - X_R'(\tau-\sigma) = \partial_\tau \tilde{X}, \tag{18.14}$$

and similarly

$$\partial_\tau X = X_L'(\tau+\sigma) + X_R'(\tau-\sigma) = \partial_\sigma \tilde{X}. \tag{18.15}$$

This makes it clear that N and D boundary conditions are exchanged by T-duality. Indeed, X is of NN type, and $\tilde{X}$ is of DD type. We summarize our facts on open string T-duality by

$$\begin{aligned} &X = X_L + X_R\,, && \tilde{X} = X_L - X_R, \\ &\partial_\sigma X = \partial_\tau \tilde{X}, && \partial_\tau X = \partial_\sigma \tilde{X}. \end{aligned} \tag{18.16}$$

The differential relations in (18.1) can be used to prove again, more conceptually, that in the dual spacetime the open string winds around the compactified dimension:

$$\tilde{X}(\tau,\pi) - \tilde{X}(\tau,0) = \int_0^\pi d\sigma\,\partial_\sigma \tilde{X} = \int_0^\pi d\sigma\,\partial_\tau X = 2\pi\alpha' \int_0^\pi d\sigma\,\mathcal{P}^\tau = 2\pi\alpha' p,$$

in agreement with (18.13).

The present considerations extend trivially to cases in which more than one dimension is compactified. Consider a D25-brane in a world where k spatial dimensions are curled up into circles. A simultaneous T-duality transformation on each circle gives a physically equivalent world where we have a D$(25-k)$-brane and each circle is replaced by a circle with the dual radius. There is also no need to start with a D25-brane. If we curl up one circle only and wrap one direction of a D3-brane around it, a T-duality along the circle will give a D2-brane on a spacetime with a circle of dual radius. In general, if a Dp-brane stretches around a compact dimension, T-duality along this direction will give a D$(p-1)$-brane at some fixed point on a circle of dual radius. All of these results hold trivially because a T-duality along a given direction does not affect the open string coordinates corresponding to other directions.

Since we use the light-cone gauge throughout, the T-duality that takes a Dp-brane into a D$(p-1)$-brane has only been established for $p \geq 2$. Indeed, two or more spatial coordinates must be Neumann: X^1, to form together with X^0 the light-cone coordinates $X^\pm$, and X along the compact direction. It is nevertheless true that T-duality holds for $p = 1$. If we have a D1-brane wrapped around a circle, the equivalent T-dual configuration has a D0-brane at some point on a circle of dual radius. This result can be proven using covariant quantization of open strings.

18.2 $U(1)$ gauge transformations

In this section we will examine Maxwell gauge transformations in detail. Gauge field configurations related by gauge transformations are physically equivalent. It is therefore necessary to understand gauge transformations in order to find the possible inequivalent gauge field configurations. This is interesting because, just as D-branes change type under T-duality, gauge field configurations on a D-brane also change in a dramatic way under T-duality. Our previous understanding of gauge transformations does not suffice because of two complicating factors. First, we will deal with compact dimensions, where topological effects become important. Second, we include the effects of charges. Indeed, we learned in Section 16.3 that an open string endpoint lying on a D-brane is seen as a charged particle by the Maxwell field that lives on the D-brane.

The analysis of gauge transformations in the presence of charges is easily done by considering the Schrödinger equation for a nonrelativistic charged particle. In natural units ($\hbar = c = 1$), the Hamiltonian for a particle with mass m and charge q (Problem 5.4) is

$$H = \frac{1}{2m}(\vec{p} - q\vec{A})^2 + q\Phi. \tag{18.17}$$

Recall that the potentials have natural units of mass, or inverse length, and that the charge q is dimensionless. The Schrödinger equation is then

$$i\frac{\partial\psi}{\partial t} = \frac{1}{2m}\left(\frac{\nabla}{i} - q\vec{A}\right)^2\psi + q\Phi\psi. \tag{18.18}$$

To test the gauge invariance of the classical mechanics of a charged point particle (Section 16.3), we only had to vary the electromagnetic potentials A_μ. The Schrödinger equation, however, is not invariant under just a change in the vector potential. A gauge transformation in quantum mechanics involves changes in both the potentials *and* the wavefunction. The content of the Schrödinger equation remains invariant when the following changes are made simultaneously:

$$\begin{aligned} \vec{A} &\longrightarrow \vec{A}' = \vec{A} + \nabla\chi, \\ \Phi &\longrightarrow \Phi' = \Phi - \frac{\partial\chi}{\partial t}, \\ \psi &\longrightarrow \psi' = \exp(iq\chi)\ \psi. \end{aligned} \tag{18.19}$$

Here $\chi(x)$ is a function of spacetime. One can readily show (Problem 18.1) that the Schrödinger equation for the primed variables,

$$i\frac{\partial\psi'}{\partial t} = \frac{1}{2m}\left(\frac{\nabla}{i} - q\vec{A}'\right)^2\psi' + q\Phi'\psi', \tag{18.20}$$

is equivalent to the original Schrödinger equation (18.18). For future reference, let $U(x)$ denote the phase factor in the new wavefunction:

$$U(x) \equiv \exp(iq\chi(x)), \quad \psi' = U\,\psi. \tag{18.21}$$

Now comes the key shift in viewpoint. In the past (and in the first two lines of (18.19), as well!), we have always written the vector potential gauge transformations as

$$A'_\mu = A_\mu + \partial_\mu\chi, \tag{18.22}$$

thinking of $\chi(x)$ as the gauge parameter. We will see that, in fact,

the gauge parameter is $U(x)$. (18.23)

While this change is sometimes inconsequential, it has effects when there are compact dimensions. If U is the gauge parameter, we must be able to write the gauge transformation of A_μ in terms of U. The object $(\partial_\mu U)U^{-1}$ is useful for this purpose, because

$$(\partial_\mu U)U^{-1} = \big(\partial_\mu \exp(iq\chi)\big)\exp(-iq\chi) = iq\,\partial_\mu\chi. \tag{18.24}$$

It then follows that (18.22) is reproduced by

$$A'_\mu = A_\mu - \frac{i}{q}(\partial_\mu U)U^{-1}. \tag{18.25}$$

Maxwell theory is called a $U(1)$ gauge theory, because the gauge parameter $U(x)$ can be viewed, for any fixed x, as an element of the group $U(1)$. To understand this we must explain what the $U(1)$ group is and why it is relevant to gauge transformations.

The group $U(1)$ is literally defined as the group of unitary matrices of size 1-by-1. A 1-by-1 matrix has just one entry u, and the unitarity condition, stating that the Hermitian conjugate matrix equals the inverse matrix, gives $u^*u = 1$. We conclude that u is a complex

number of unit norm: a phase factor $u = \exp(i\theta)$. The set of complex numbers of unit norm forms a group under multiplication: the multiplication of two complex numbers of unit norm gives a complex number of unit norm, multiplication is associative, there is an identity element $u = 1$ in the group, and any number $\exp(i\theta)$ in the group has an inverse $\exp(-i\theta)$ in the group.

For any fixed spacetime point x, the gauge parameter $U(x)$ is a phase factor; the possible values of $U(x)$ are therefore in one-to-one correspondence with the elements of the group $U(1)$. Since U is spacetime dependent, we have, in fact, a $U(1)$-group worth of possible gauge parameters at *each* spacetime point. The gauge parameter U is best thought of as a function from spacetime to the $U(1)$ group; for each spacetime point we get a $U(1)$ group element.

The concept of a group is relevant to gauge theory because gauge transformations performed in sequence combine according to the rule of group multiplication. We claim that

> a gauge transformation with parameter U_2, followed by a gauge transformation with parameter U_1, is a gauge transformation with parameter $U_1 U_2$.

This composition law is easily verified for the wavefunction ψ. Equation (18.21) implies that the sequence of gauge transformations gives

$$\psi(x) \longrightarrow U_2(x)\psi(x) \longrightarrow U_1(x)(U_2(x)\psi(x)) = (U_1U_2)(x)\,\psi(x). \tag{18.26}$$

The gauge field A_μ transforms nontrivially under a gauge transformation only if U is not a constant. The composition law also holds for the gauge field:

$$A_\mu \longrightarrow A_\mu - \frac{i}{q}(\partial_\mu U_2)U_2^{-1} \longrightarrow A_\mu - \frac{i}{q}(\partial_\mu U_2)U_2^{-1} - \frac{i}{q}(\partial_\mu U_1)U_1^{-1}, \tag{18.27}$$

since the last two terms on the right-hand side are equal to

$$-\frac{i}{q}\left(\partial_\mu\,(U_1U_2)\right)\,(U_1U_2)^{-1}. \tag{18.28}$$

18.3 Wilson lines on circles

In this section we apply our discussion of gauge transformations to a world with a compactified dimension. We will find gauge field configurations and effects that are reminiscent of the classic Aharonov–Bohm effect. In the setup for the Aharonov–Bohm effect, there is a solenoid which produces a magnetic field that is confined to its interior. Although a charged particle that is moving outside the solenoid is moving in a region with $\vec{B} = 0$, the wavefunction of the particle is affected by the vector potential $\vec{A}$ outside the solenoid. Since the solenoid produces a nonzero magnetic field, the vector potential outside the solenoid cannot vanish. In a simple gauge, for example, the vector potential goes around the solenoid. The lesson of the Aharonov–Bohm effect is that in quantum mechanics there are magnetic effects in regions of space that have zero magnetic field. This happens because the vector

potential is nontrivial. In the Aharonov–Bohm effect, the physical source of the effect is ultimately the magnetic field, since without it the interference effects would vanish. As we will soon see, if we have a compact dimension, there are physical effects from vector potentials even when the magnetic field vanishes *everywhere*.

Quick calculation 18.1 Show that, if a solenoid has magnetic flux $\Phi = \int \vec{B} \cdot d\vec{a}$, then $\oint \vec{A} \cdot d\vec{l} = \Phi$ for a closed curve that surrounds the solenoid. How is this curve oriented?

Suppose the spatial dimension x is compactified into a circle. In addition, assume that the vector potential $\vec{A}$ vanishes except for a component A_x along the circle. In this case there is no magnetic field anywhere! This may at first seem surprising, since a similar vector potential wraps around a circular solenoid's magnetic field. But because there is no space "inside" the circular dimension, there is no magnetic field. As we will see, this vector potential $\vec{A}$ is not a gauge artifact; it affects the physics of particles just as if it were the consequence of a magnetic field. In string theory, the analysis of such a setup is necessary to understand how T-duality works for configurations of D-branes that do not coincide.

Consider a nonzero A_x that is a constant, independent of all coordinates, including the compact dimension x. Can such a field exist? A field configuration can exist if it satisfies the field equations of the theory. A constant A_x, with all other components of A_μ equal to zero, gives zero $F_{\mu\nu}$, and this satisfies the source-free Maxwell equations. This solution can be understood intuitively in three spatial dimensions. Consider a circular solenoid of radius r_0, whose axis coincides with the z axis. If the solenoid flux is $\Phi > 0$, the vector potential can be chosen to be in the azimuthal direction and to have a magnitude $\Phi/(2\pi\rho)$, where ρ is the distance to the axis of the solenoid. This configuration satisfies all equations of motion outside the solenoid. Imagine now selecting a thin cylinder $R < \rho < R + \epsilon$, where $R > r_0$ and ϵ is infinitesimal. The equations of motion are satisfied on this thin cylinder. The solution that represents a constant $\vec{A}$ which wraps around a compact dimension is obtained by throwing away all the space inside and outside the cylinder and letting $\epsilon \to 0$.

Similar reasoning explains why a constant electric field can wrap around a compact dimension. It is possible to have

$$\oint_\Gamma \vec{E} \cdot d\vec{l} \neq 0 \tag{18.29}$$

if Γ wraps around a compact dimension. Such an electrostatic field is not allowed in ordinary space. For any closed curve Γ in ordinary space there is a surface S whose boundary is Γ. Using Stokes' theorem and the Maxwell equation $\nabla \times \vec{E} = 0$, we then have

$$\oint_\Gamma \vec{E} \cdot d\vec{l} = \int_S (\nabla \times \vec{E}) \cdot d\vec{a} = 0. \tag{18.30}$$

If Γ wraps around a compact dimension, however, there is no surface for which Γ is the boundary, and (18.30) then does not apply. Again, the existence of gauge potentials that satisfy all relevant equations is the trustworthy guide. Since $\vec{E} = -\partial \vec{A}/\partial t$, a constant E_x in the x direction can be obtained with $A_x = -E_x t$. This gauge field satisfies all the relevant field equations.

So let us consider a vector potential $A_x(x)$ along the compact dimension x. We will try to classify the physically inequivalent configurations. Under a gauge transformation A_x changes as (18.19)

$$A_x \longrightarrow A_x + \frac{\partial \chi}{\partial x}. \tag{18.31}$$

One might think that, since x and $x + 2\pi R$ represent the same point, the parameter χ should satisfy

$$\chi(x + 2\pi R) \stackrel{?}{=} \chi(x). \tag{18.32}$$

To understand the implications of this candidate condition, we examine the line integral of the vector potential around the circle:

$$w \equiv q \oint dx\ A_x. \tag{18.33}$$

Here we have defined the dimensionless quantity w associated with the gauge field along the circle. We also define the *holonomy* W of the gauge field:

$$W \equiv \exp(iw) = \exp\Big(iq \oint dx\ A_x\Big). \tag{18.34}$$

W is called a *Wilson line*. In our present context, the Wilson line associated with a closed curve is simply the calculable phase factor that depends on the values of the gauge field along the curve. After the gauge transformation (18.31) w is changed into w':

$$w' = q \oint dx \Big(A_x + \frac{\partial \chi}{\partial x}\Big) = w + q\big(\chi(x_0 + 2\pi R) - \chi(x_0)\big), \tag{18.35}$$

where we used some arbitrary reference point x_0 on the circle to do the integral. If we assume the periodicity condition (18.32) then $w' = w$. We will soon see that this is not the correct picture.

Our analysis is flawed because of a failure of imagination, rather than a technical mistake. Equation (18.32) is sensible if $\chi(x)$ is the fundamental gauge parameter. But we claimed in Section 18.2 that the gauge parameter is $U(x)$. If U is the gauge parameter, it must be periodic on the circle:

$$U(x + 2\pi R) = U(x). \tag{18.36}$$

Since $U = \exp(iq\chi)$, the above equation implies

$$q\,\chi(x + 2\pi R) = q\,\chi(x) + 2\pi m, \quad m \in \mathbb{Z}, \tag{18.37}$$

or, equivalently,

$$q\,\big(\chi(x + 2\pi R) - \chi(x)\big) = 2\pi m. \tag{18.38}$$

Note the dramatic change from (18.32). If U is the gauge parameter, $\chi(x)$ need only be quasi-periodic! Moreover, looking back at equation (18.35), we see that

$$w' = w + 2\pi m. \tag{18.39}$$

Therefore the physics does not change when we replace w by $w + 2\pi m$. We write $w \sim w + 2\pi m$, or

$$q \oint dx\ A_x \sim q \oint dx\ A_x + 2\pi m. \tag{18.40}$$

As before, we interpret the above equation to mean that w lives on a unit circle, or that all physically inequivalent values of w are represented on the fundamental domain

$$w = q \oint dx\ A_x \in [0, 2\pi). \tag{18.41}$$

The most natural interpretation is that w is really an angle. Thus we will write

$$\theta \equiv w = q \oint dx\ A_x. \tag{18.42}$$

This is an important result. In the presence of compact dimensions, line integrals of the vector potential are angular variables. Gauge transformations act as $\theta \to \theta + 2\pi m$. We will see that in string theory the abstract angle θ has a concrete physical interpretation. It is worth noting that the Wilson line $W = \exp(i\theta)$ is gauge invariant. Gauge equivalent θ give the same holonomy W.

It is possible to write explicitly χ that are not single valued, but that, nevertheless, lead to a single-valued U because they obey (18.38). There are infinitely many physically equivalent choices, but the simplest one takes χ to be linear in the compact coordinate x:

$$q\,\chi = (2\pi m)\,\frac{x}{2\pi R} = \frac{mx}{R}. \tag{18.43}$$

As x changes from 0 to $2\pi R$, the quantity $q\chi$ changes by $2\pi m$, as required in (18.38). For this choice of χ, the gauge transformation changes the gauge field by a constant:

$$q\,A_x(x) \longrightarrow q\,A_x(x) + \frac{m}{R}. \tag{18.44}$$

This gauge transformation implies that we can identify

$$q\,A_x \sim q\,A_x + \frac{m}{R}. \tag{18.45}$$

The natural realization of a Wilson line is with *constant* A_x. In this case, (18.42) gives

$$q\,A_x = \frac{\theta}{2\pi R}. \tag{18.46}$$

The presence of a Wilson line on a circle changes significantly the physics of a charged particle. To see how this happens, we examine the Schrödinger equation

$$\frac{1}{2m}\left(\frac{1}{i}\frac{\partial}{\partial x} - q\,A_x\right)^2 \psi = i\frac{\partial \psi}{\partial t}. \tag{18.47}$$

With A_x given in (18.46), energy eigenstates $\psi(x,t) = e^{-iEt}\psi(x)$ satisfy

$$\frac{1}{2m}\left(\frac{1}{i}\frac{\partial}{\partial x} - \frac{\theta}{2\pi R}\right)^2 \psi(x) = E\psi(x). \tag{18.48}$$

The wavefunction $\psi(x)$ must be periodic: $\psi(x + 2\pi R) = \psi(x)$. Therefore the solutions take the form $\psi_\ell(x) \sim \exp(i\ell x/R)$, with $\ell \in \mathbb{Z}$. The corresponding energy levels are

$$E_\ell = \frac{1}{2m}\left(\frac{\ell}{R} - \frac{\theta}{2\pi R}\right)^2. \tag{18.49}$$

Note that the energy levels are shifted if $\theta \neq 0$. In particular, the $\theta = 0$ degeneracy between $\pm\ell$ energy levels is broken. Since θ is an angle variable the energy levels must remain unchanged when we let $\theta \to \theta + 2\pi$. Since ℓ varies over all $\mathbb{Z}$, we can easily see that the *set* of energy levels is unchanged under this transformation. The shift $\theta \to \theta + 2\pi$ is compensated by letting $\ell \to \ell + 1$. This confirmation that θ is naturally an angle variable gives credence to our claim that $U(x)$ is the fundamental gauge parameter. If χ had been the gauge parameter all values of θ would be gauge inequivalent.

Quick calculation 18.2 Consider a gauge transformation with χ linear in x that takes $\theta \to \theta + 2\pi$. What does it do to ψ_ℓ?

18.4 Open strings and Wilson lines

We are now finally in a position to study the physics of open strings on D-branes that have gauge fields characterized by holonomies. T-duality will provide a physical interpretation for the angle variable that represents the gauge-field holonomy.

Consider a Dp-brane that wraps around a compact dimension x. This simply means that one of the spatial directions along the Dp-brane is the x direction. A gauge field lives on the world-volume of the Dp-brane. Let us now assume that this gauge field is such that $q\oint A_x dx$ takes the value θ. What is the T-dual picture of this brane configuration? We learned before that the dual world has a D$(p-1)$-brane located at some position on a circle of dual radius. But what does the parameter θ correspond to? The most obvious guess is correct: θ parameterizes the position of the D$(p-1)$-brane on the dual circle! This gives a very concrete meaning to the formerly abstract angle variable. The situation is illustrated in Figure 18.2.

What is the evidence for this interpretation of θ? First, the periodicity makes physical sense since a D$(p-1)$-brane at θ is the same as a D$(p-1)$-brane at $\theta + 2\pi$. Second, just as the position of the D$(p-1)$-brane on the circle does not affect the spectrum of open strings that end on it, the Wilson line does not affect the spectrum of open strings with endpoints on the Dp-brane. This is not readily apparent, given that charged particles are in fact affected by Wilson lines. In the case of strings, however, the endpoints have opposite charges so the string as a whole is neutral and the Wilson line has no effect. If a D-brane wraps the compact dimension, the mass-squared for open string states is given as

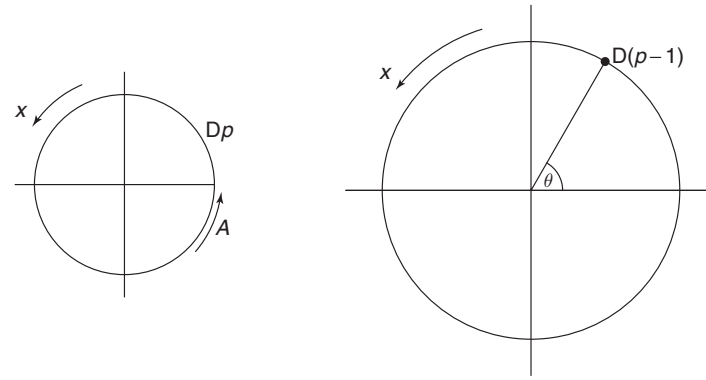

Fig. 18.2 A Dp-brane wrapped on a circle, with a vector potential along the circle. In the T-dual picture, the line integral of this gauge field becomes the angle that parameterizes the position of the D$(p-1)$-brane on the dual circle.

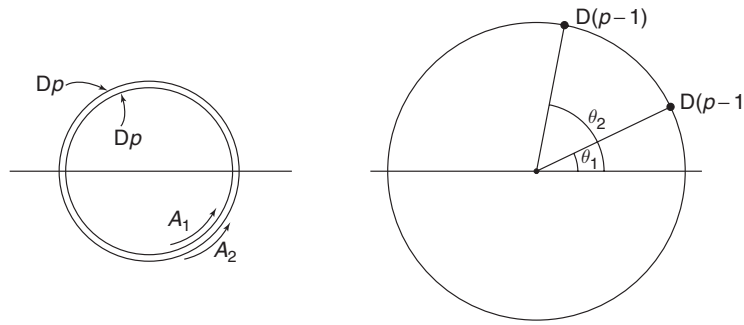

Fig. 18.3 Two Dp-branes that wrap on a circle, one with parameter θ_1 and the other with parameter θ_2. In the T-dual picture, we have two D$(p-1)$-branes separated by an angle $\theta_2 - \theta_1$.

$$M^2 = p^2 + \frac{1}{\alpha'}(N^\perp - 1), \quad p = \frac{\ell}{R}, \tag{18.50}$$

where p is the quantized momentum in the compact direction, and $N^\perp$ is the appropriate number operator. For a particle, the addition of a Wilson line resulted in p changing into $p - qA$ in the Hamiltonian. This implied letting

$$\frac{\ell}{R} \to \frac{\ell}{R} - \frac{\theta}{2\pi R}, \tag{18.51}$$

as we saw in (18.49). What happens with strings? Since strings have two endpoints with opposite charges, if both endpoints lie on the same Dp-brane the effect is cancelled: $p \to p - qA + qA = p$. To give direct evidence for the interpretation of θ we need more than one D$(p-1)$-brane on the circle. In this case, the relative positions of the branes matter.

Consider therefore a string stretching between the two Dp-branes. The situation is shown in Figure 18.3. Each D-brane has its own Maxwell field. Suppose the negatively

charged endpoint lies on the first Dp-brane, with Wilson line parameter θ_1, and the positively charged endpoint lies on the second Dp-brane, with Wilson line parameter θ_2. The momentum is then shifted from p to $p + qA_1 - qA_2$, and therefore

$$\frac{\ell}{R} \to \frac{\ell}{R} - \frac{\theta_2}{2\pi R} + \frac{\theta_1}{2\pi R}. \tag{18.52}$$

As a result, the mass-squared formula reads

$$M^2 = \left(\frac{2\pi\ell - (\theta_2 - \theta_1)}{2\pi R}\right)^2 + \frac{1}{\alpha'}(N^\perp - 1), \quad \ell \in \mathbb{Z}. \tag{18.53}$$

If $\theta_1 = \theta_2$ the effects of the holonomies cancel out. As we argued before, the same would happen if both endpoints of the string were on the same brane.

The configuration T-dual to the two Dp-branes with different θ parameters consists of two D$(p-1)$-branes with different positions, corresponding to the two values of θ. We can easily verify that the mass formulae are consistent. For example, consider the $\ell = 0$ states of the string that stretches between the two Dp-branes:

$$M^2 = \left(\frac{\theta_2 - \theta_1}{2\pi R}\right)^2 + \frac{1}{\alpha'}(N^\perp - 1). \tag{18.54}$$

If, as we claim, θ_1 and θ_2 are physical angles, then in the dual world there must be two D$(p-1)$-branes such that the angular difference between them is $\theta_2 - \theta_1$. The mass-squared for a string that is stretched between these branes receives a contribution equal to the length of the string multiplied by its tension, squared:

$$\begin{aligned} M^2 &= \left((\theta_2 - \theta_1)\tilde{R}T_0\right)^2 + \frac{1}{\alpha'}(N^\perp - 1) \\ &= \left((\theta_2 - \theta_1)\frac{\alpha'}{R}\frac{1}{2\pi\alpha'}\right)^2 + \frac{1}{\alpha'}(N^\perp - 1) \\ &= \left(\frac{\theta_2 - \theta_1}{2\pi R}\right)^2 + \frac{1}{\alpha'}(N^\perp - 1), \end{aligned} \tag{18.55}$$

which exactly matches (18.54). This is strong evidence for our physical interpretation of the θ parameter (18.42) as the angular position of D-branes in the T-dual configuration.

Two important points about T-duality are examined in the problems. The first concerns the fact that the string coupling constant must change under a T-duality transformation along a circle. We learned in Section 3.9 that Newton's constant in the effective lower-dimensional world is related to the higher-dimensional Newton constant and the volume of the extra dimensions. This volume changes under T-duality since one circle changes its radius. Moreover, the higher-dimensional Newton constant is set by the value of the string coupling and α' (see Section 13.4). If T-duality is a symmetry of the theory, then the lower-dimensional Newton constant must be unchanged. After all, this constant is observable. If the string coupling is g and the radius of the circle to be dualized is R, after T-duality the value $\tilde{g}$ of the string coupling is (Problem 18.5)

$$\tilde{g} = \frac{\sqrt{\alpha'}}{R}\, g. \tag{18.56}$$

The second point deals with the tension of D-branes (Problem 18.6). For a static string, the tension multiplied by the length gives the mass. D-branes also have tension and mass. For a static Dp-brane, the tension T_p of the brane multiplied by its volume V_p gives the mass of the brane. Our work with D-branes has assumed that they are fixed hyperplanes, or very heavy objects that are hardly affected by the open strings that are attached to them. In fact, the tension T_p of a D-brane goes to infinity as the string coupling $g \to 0$. The fact that the mass of a wrapped D-brane must be unchanged under T-duality will allow you to prove that

$$T_p(g) = \frac{\tau_p}{g} \quad \text{and} \quad T_{p-1}(g) = T_p(g) \cdot 2\pi \sqrt{\alpha'}. \tag{18.57}$$

Here τ_p is a p-dependent, but g-independent constant that includes a suitable factor of α' to give the correct units to the tension. The second relation implies that the tensions of all D-branes are related. It is conventional to choose the precise definition of the string coupling (see (13.85)) in such a way that the tension of the D1-brane is given by

$$T_1(g) = \frac{1}{2\pi\alpha'} \frac{1}{g}. \tag{18.58}$$

This formula is reminiscent of the string tension $1/(2\pi\alpha')$, which lacks the factor $1/g$. With the help of (18.57) the tensions of all D-branes are now determined. This formula also gives the correct value for a D0-brane, whose tension is simply its mass.

Quick calculation 18.3 Find the explicit p-dependent expression of $T_p(g)$. Confirm that it has the correct units.

Problems

Problem 18.1 Gauge invariance of the Schrödinger equation.

To prove the gauge invariance of the Schrödinger equation it is useful to note that, for any function M, and for U defined as in (18.21), we have

$$\Big(\frac{\nabla}{i} - q\vec{A} - q\nabla\chi\Big) UM = U\Big(\frac{\nabla}{i} - q\vec{A}\Big) M.$$

Prove this result and use it to show that (18.20) is equivalent to (18.18).

Problem 18.2 Explicit T-duality of DN string coordinates.

Consider the expansion (15.72) for an ND string coordinate. Find the separate left-moving and right-moving pieces, construct the dual string coordinate, and verify that it is of DN type.

Problem 18.3 T-duality invariance of the Hamiltonian.

Find the relations between $\dot{X} \pm X'$ and $\dot{\tilde{X}} \pm \tilde{X}'$. Use these relations, together with (15.8), to explain why the string Hamiltonian is unchanged under T-duality.

Problem 18.4 T-duality of a Dp-brane with an electric field.

Assume we have a constant electric field E_x on the world-volume of a Dp-brane that is wrapped around a circle of radius R in the x direction. Give the time-dependent A_x that corresponds to this electric field. Using the T-dual of a Dp-brane with a holonomy, show that the T-dual of the Dp-brane with an electric field is a D$(p-1)$-brane on a circle of dual radius, moving along the circle with a velocity $v_x = 2\pi\alpha' E_x$. As an additional consistency check, use the original world of the Dp-brane with the electric field to calculate how long the D$(p-1)$-brane takes to go around the dual circle.

Problem 18.5 T-duality and the string coupling.

Consider a D-dimensional spacetime where p spatial dimensions have been curled up. One of them is a circle of radius R, and the rest form a space of volume V_{p-1}. Let $\hat{G}$ denote Newton's constant in the $(D-p)$-dimensional effective spacetime. Write an expression for $\hat{G}$ in terms of the string coupling g, α', R, and V_{p-1}. Imagine now performing a T-duality along the circle of radius R. Show that the invariance of $\hat{G}$ requires changing the string coupling g into $\tilde{g}$, as given in (18.56).

Problem 18.6 D-brane tension and descent relations.

The tension T_p of a D-brane can be written in the form $T_p = \tau_p\, h(g)$, where g is the string coupling, h is a function to be determined below, and τ_p is a p-dependent constant that includes some suitable power of α'. Consider the setup of Problem 18.5, and imagine wrapping a Dp-brane around the compact dimensions. The mass of this object, perceived as a point particle by the lower-dimensional observer, must be unchanged under T-duality. Use this condition to show that

$$\tau_p\, h(g)\, 2\pi R = \tau_{p-1}\, h(g\sqrt{\alpha'}/R).$$

Use this equation to prove (18.57). The first relation in (18.57) requires absorbing a p-independent constant into the definition of τ_p.

19 Electromagnetic fields on D-branes

We now begin a study of D-branes that carry electric or magnetic fields on their world-volume. Open strings couple to these electromagnetic fields at their endpoints. Using the tools of T-duality we show that a D-brane with an electric field is physically equivalent to a moving D-brane with no electric field. The constraint that a D-brane cannot move faster than light implies that the strength of an electric field cannot exceed a certain maximum value. We also show that a Dp-brane with a magnetic field is T-duality equivalent to a tilted D$(p-1)$-brane with no magnetic field. Alternatively, the magnetic field on the Dp-brane can be thought of as being created by a distribution of dissolved D$(p-2)$-branes.

19.1 Maxwell fields coupling to open strings

Among the quantum states of open strings attached to a D-brane we found photon states with polarizations and momentum along the D-brane directions. We thus deduced that a Maxwell field lives on the world-volume of a D-brane. The existence of this Maxwell field was in fact necessary to preserve the gauge invariance of the term that couples the Kalb–Ramond field to the string in the presence of a D-brane. We also learned that the endpoints of open strings carry Maxwell charge.

Since any D-brane has a Maxwell field, it is physically reasonable to expect that background electromagnetic fields can exist: there may be electric or magnetic fields that permeate the D-brane. If the universe was the world-volume of a D-brane, an intergalactic magnetic field would be an example of a background field. It is customary to think of backgrounds as solutions of classical field equations. Historically, electromagnetism was first studied using familiar backgrounds, such as the magnetic field of the earth, or the static electric field of a charge. The study of these and other backgrounds led to a set of classical equations, Maxwell's equations, which gave these backgrounds as classical solutions. Eventually, physicists developed a quantum theory of electromagnetism, and this theory predicts photon states. Note that our string theory discovery of electromagnetism ran completely in reverse. We found string quantum states that could be identified as photon states, and through an analysis of the Schrödinger equations satisfied by the quantum states, we recovered the classical equations of electromagnetism. Now, we want to study background electromagnetic fields!

Candidate background fields can be tested to see if they are consistent with string theory. As we first discussed at the end of Section 16.1, to do so, we must re-quantize the string, taking into account the effects of the candidate background fields. If the quantization is successful then we may conclude that the candidate background fields are permitted in string theory. After all, a successful quantization establishes that quantum string states can consistently propagate on the given background. Although the quantizations we performed previously only applied to strings in the absence of nontrivial background fields, we already studied the effects of at least one background field. Our quantization of closed strings gave rise to Kalb–Ramond states, and as closed strings are not confined to D-branes, we concluded that Kalb–Ramond fields live on the full spacetime. Our previous discussion was therefore an analysis of nontrivial Kalb–Ramond backgrounds. In that discussion we introduced into the string action a new term, which couples the string to the Kalb–Ramond background field (see (16.3)). We did not re-quantize the closed string with this new term, except in the simple case considered in Problem 17.4. Rather, we examined the implications of the new coupling for string motion and gauge invariance.

In this chapter we study the effects of electromagnetic backgrounds on open strings. We will not embark on the details of the quantization; instead, we will assume that the quantization works out properly (as it does, for the backgrounds we will consider). In the present section we derive the equations of motion for open strings in the presence of background electromagnetic fields, and later we will use the tools of T-duality to gain new physical insights. The discussion of electromagnetic fields on D-branes continues in Chapter 20, where we will show that their dynamics is governed by the Born–Infeld theory of nonlinear electrodynamics.

We learned in Chapter 16 how to describe the coupling of Maxwell fields to strings. The string endpoints couple to the Maxwell potential A_m in the same way as a charged particle does. The coupling terms were given in equation (16.54); adding them to the string action gives

$$S = \int d\tau d\sigma\, \mathcal{L}(\dot{X}, X') + \int d\tau\, A_m(X) \frac{dX^m}{d\tau}\Big|_{\sigma=\pi} - \int d\tau\, A_m(X) \frac{dX^m}{d\tau}\Big|_{\sigma=0}. \tag{19.1}$$

Here $\mathcal{L}$ denotes the Nambu–Goto Lagrangian density. In our index convention $\mu, \nu, \ldots$ are spacetime indices, which run from 0 to d, while $m, n, \ldots$ are brane world-indices, which run from 0 to p. The index on the gauge potential A_m, for example, is a brane world-index. The indices $i, j, \ldots$ are spatial indices on the brane, which run from 1 to p, and $a, b, \ldots$ are indices for directions normal to the brane, which run from $p+1$ to d. We will only consider backgrounds for which the electromagnetic field strength F_{mn} is a constant. If the only nonvanishing entries are the $F_{0i} (= -F_{i0})$ then the background is purely electric. If the only nonvanishing entries are the F_{ij} then the background is purely magnetic. For a constant F_{mn} the gauge potentials can be chosen to be

$$A_n(x) = \frac{1}{2} F_{mn}\, x^m. \tag{19.2}$$

Quick calculation 19.1 Confirm that $\partial_m A_n - \partial_n A_m$ is equal to F_{mn}.

Making use of (19.2) the action S becomes

$$S = \int d\tau d\sigma\, \mathcal{L}(\dot{X}, X') + \frac{1}{2}\int d\tau\, F_{mn}\Big(X^m \partial_\tau X^n\,|_{\sigma=\pi} - X^m \partial_\tau X^n\,|_{\sigma=0}\Big). \tag{19.3}$$

To find the equations of motion we use our earlier notation where $\mathcal{P}^\tau_\mu = \partial\mathcal{L}/\partial\dot{X}^\mu$ and $\mathcal{P}^\sigma_\mu = \partial\mathcal{L}/\partial X^{\mu\prime}$. Note that $\mathcal{P}^\tau_\mu$ is not the complete momentum conjugate to X^μ, since the present $\mathcal{L}$ does not include the contributions from the endpoints and is therefore not the complete Lagrangian density. Since the two endpoints enter into the string action almost symmetrically, we calculate the variation of the action by focusing only on the $\sigma = \pi$ endpoint:

$$\begin{aligned}\delta S = &\int d\tau d\sigma \left(\mathcal{P}^\tau_\mu \partial_\tau \delta X^\mu + \mathcal{P}^\sigma_\mu \partial_\sigma \delta X^\mu\right)\\ &+ \frac{1}{2}\int d\tau F_{mn}\Big(\delta X^m \partial_\tau X^n + X^m \partial_\tau \delta X^n\Big)\Big|_{\sigma=\pi} + \cdots,\end{aligned} \tag{19.4}$$

where the dots indicate terms contributed by the $\sigma = 0$ endpoint. The wave equation $\partial_\tau \mathcal{P}^\tau_\mu + \partial_\sigma \mathcal{P}^\sigma_\mu = 0$ continues to apply, but at the string endpoints there is a new constraint. The only effect of electromagnetic fields is a change in the boundary conditions. We can determine the boundary conditions by paying attention to the terms that involve variations at the string endpoints. Since total τ derivatives are not relevant to the equations of motion, the boundary terms that we should examine are

$$\delta S = \int d\tau d\sigma\, \partial_\sigma(\mathcal{P}^\sigma_\mu \delta X^\mu) + \int d\tau \delta X^m F_{mn} \partial_\tau X^n|_{\sigma=\pi} + \cdots. \tag{19.5}$$

For the coordinates normal to the brane, we have the usual Dirichlet boundary condition $\delta X^a = 0$. Focusing now on the variations δX^m along the brane, we find

$$\delta S = \int d\tau\, \delta X^m \Big(\mathcal{P}^\sigma_m + F_{mn}\partial_\tau X^n\Big)\Big|_{\sigma=\pi} + \cdots. \tag{19.6}$$

Since the X^m are coordinates on the brane, they carry free boundary conditions: no constraint can be imposed on the variations δX^m. As a result,

$$\mathcal{P}^\sigma_m + F_{mn}\partial_\tau X^n = 0, \quad \text{for} \quad \sigma = 0, \pi. \tag{19.7}$$

Quick calculation 19.2 Convince yourself that (19.7) applies for $\sigma = 0$.

To simplify the boundary conditions we must choose a particular gauge. We impose the orthonormality conditions

$$\dot{X}\cdot X' = 0, \quad \dot{X}^2 + X'^2 = 0, \tag{19.8}$$

which hold for both the static and light-cone gauges. We then have

$$\mathcal{P}^\sigma_\mu = -\frac{1}{2\pi\alpha'}\partial_\sigma X_\mu, \tag{19.9}$$

and the boundary condition becomes

$$\partial_\sigma X_m - 2\pi\alpha' F_{mn}\partial_\tau X^n = 0, \quad \sigma = 0, \pi. \tag{19.10}$$

Let us look briefly at the physics that is contained in this boundary condition in the case of a pure magnetic field. Since $F_{0i} = 0$, there is no change in the boundary condition for X^0 – it is still Neumann. On the other hand, for spatial directions along which there is a magnetic field we get

$$\partial_\sigma X_i - 2\pi\alpha' F_{ij}\partial_\tau X^j = 0, \quad \sigma = 0, \pi. \tag{19.11}$$

This boundary condition is of mixed type; it is neither Neumann nor Dirichlet. Suppose that the only nonvanishing component of the magnetic field is $F_{23} = -F_{32} \equiv B$. The coordinates X_1 and X_i with $i > 3$ then satisfy Neumann boundary conditions, but X_2 and X_3 satisfy

$$\begin{aligned} \partial_\sigma X_2 - 2\pi\alpha' B\, \partial_\tau X^3 = 0, \\ \partial_\sigma X_3 + 2\pi\alpha' B\, \partial_\tau X^2 = 0. \end{aligned} \tag{19.12}$$

If B is very large, then the first term in each equation is negligible when compared to the second term. In this case the boundary conditions become, approximately,

$$\partial_\tau X^2 = \partial_\tau X^3 = 0, \quad \sigma = 0, \pi. \tag{19.13}$$

As $F_{23} = B$ becomes infinitely large, the motion of the string endpoints along the directions x^2 and x^3 of the brane is frozen! The string coordinates X^2 and X^3 have become Dirichlet. It is as if the original Dp-brane is now filled with infinitely many D$(p-2)$-branes, one at each possible value of (x^2, x^3). The strings end on those D$(p-2)$-branes and cannot change their positions in the (x^2, x^3)-plane. If we begin with a D2-brane, the motion of the string endpoints along the brane will be completely frozen. Although this seems like a rather bizarre interpretation of the boundary condition, we will see later that there is quantitative truth behind it.

19.2 D-branes with electric fields

In this section we will consider a D-brane with an electric field on its world-volume. We will not examine the classical solutions that describe the motion of an open string in this background (see, however, Problem 19.2); instead, our discussion will use T-duality to learn about the physics of electric fields. We will assume that the electric field is constant and points along a compact direction on the world-volume of a Dp-brane, and we will use T-duality to relate this configuration to one in which a D$(p-1)$-brane is moving along the dual circle. This result was anticipated in Problem 18.4, but our analysis here will be more general. The constraint that the velocity of the D$(p-1)$-brane cannot exceed the velocity

of light will imply that the value of the electric field in the Dp-brane cannot exceed a critical value – electric fields on D-branes are bounded.

Our analysis will proceed in three steps. Our first step will be to find a user-friendly way of writing the boundary conditions at the string endpoints. We will then use the same language to express the equations that relate T-dual coordinates. Finally, applying a boost and a T-duality transformation we will prove the above equivalence.

Consider a Dp-brane that wraps around a compact dimension x^{25} of radius R, and assume that the brane carries an electric field along this direction:

$$F_{25,0} = E_{25} \equiv E. \tag{19.14}$$

Let us look at the string boundary conditions. The only interesting directions are X^0 and X^{25}, and from (19.10) we see that

$$\begin{aligned} \partial_\sigma X_0 - 2\pi\alpha' F_{0,25}\partial_\tau X^{25} &= 0, \\ \partial_\sigma X_{25} - 2\pi\alpha' F_{25,0}\partial_\tau X^0 &= 0. \end{aligned} \tag{19.15}$$

Using (19.14), $X_0 = -X^0$, and writing $X^{25} \equiv X$, we find

$$\begin{aligned} \partial_\sigma X^0 - \mathcal{E}\,\partial_\tau X &= 0, \\ \partial_\sigma X - \mathcal{E}\,\partial_\tau X^0 &= 0, \end{aligned} \tag{19.16}$$

where we define the dimensionless electric field $\mathcal{E}$ as

$$\mathcal{E} \equiv 2\pi\alpha' E. \tag{19.17}$$

Our next objective is to write the above boundary conditions in a more manageable form. We fit the two dynamical variables X^0 and X into a column vector and look for boundary conditions that take the form of *invertible linear relations* between derivatives of the column vector. The derivatives ∂_τ and ∂_σ are not appropriate for this. A Neumann boundary condition, for example, simply requires that the σ derivative of the column vector vanishes, and this is not an invertible linear relation between σ and τ derivatives of the column vector. To obtain relations of the desired form we introduce new partial derivatives:

$$\partial_+ \equiv \tfrac{1}{2}(\partial_\tau + \partial_\sigma), \quad \partial_- \equiv \tfrac{1}{2}(\partial_\tau - \partial_\sigma). \tag{19.18}$$

Solving for ∂_τ and ∂_σ, we obtain

$$\partial_\tau = \partial_+ + \partial_-, \quad \partial_\sigma = \partial_+ - \partial_-. \tag{19.19}$$

Rewriting the boundary conditions (19.16) in terms of $\partial_\pm$ and collecting the ∂_+ and ∂_- derivatives on the left- and right-hand sides, respectively, we find

$$\begin{aligned} \partial_+ X^0 - \mathcal{E}\,\partial_+ X &= \partial_- X^0 + \mathcal{E}\,\partial_- X, \\ -\mathcal{E}\,\partial_+ X^0 + \partial_+ X &= \mathcal{E}\,\partial_- X^0 + \partial_- X. \end{aligned} \tag{19.20}$$

Solving for $\partial_+ X^0$ and $\partial_+ X$ gives

$$\partial_+ \begin{pmatrix} X^0 \\ X \end{pmatrix} = \begin{pmatrix} \dfrac{1+\mathcal{E}^2}{1-\mathcal{E}^2} & \dfrac{2\mathcal{E}}{1-\mathcal{E}^2} \\ \dfrac{2\mathcal{E}}{1-\mathcal{E}^2} & \dfrac{1+\mathcal{E}^2}{1-\mathcal{E}^2} \end{pmatrix} \partial_- \begin{pmatrix} X^0 \\ X \end{pmatrix}. \tag{19.21}$$

This is the desired form of the boundary conditions. These equations hold at the endpoints $\sigma = 0, \pi$. Note that the matrix above has unit determinant. Moreover, the entries of the matrix become singular when $\mathcal{E} = \pm 1$.

Quick calculation 19.3 Prove equation (19.21).

We also need to express in this new language the Neumann and Dirichlet boundary conditions, and the T-duality relations, as well. If a pair of coordinates X^0 and X satisfy Neumann boundary conditions $\partial_\sigma X^0 = \partial_\sigma X = 0$, then in terms of $\partial_\pm$ these conditions read

$$\begin{aligned} \partial_+ X^0 &= \partial_- X^0, \\ \partial_+ X &= \partial_- X, \end{aligned} \tag{19.22}$$

or, in matrix form,

$$\partial_+ \begin{pmatrix} X^0 \\ X \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \partial_- \begin{pmatrix} X^0 \\ X \end{pmatrix} \quad \text{for} \quad \{X^0, X\} = \{\mathrm{N}, \mathrm{N}\}. \tag{19.23}$$

Thus, in terms of our linear relations, Neumann boundary conditions lead to an identity matrix. If X^0 is of Neumann type and $\tilde{X}$ is of Dirichlet type, then $\partial_\sigma X^0 = \partial_\tau \tilde{X} = 0$. In terms of ∂_+ and ∂_-, then,

$$\begin{aligned} \partial_+ X^0 &= \partial_- X^0, \\ \partial_+ \tilde{X} &= -\partial_- \tilde{X}, \end{aligned} \tag{19.24}$$

or, in matrix form,

$$\partial_+ \begin{pmatrix} X^0 \\ \tilde{X} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \partial_- \begin{pmatrix} X^0 \\ \tilde{X} \end{pmatrix} \quad \text{for} \quad \{X^0, \tilde{X}\} = \{\mathrm{N}, \mathrm{D}\}. \tag{19.25}$$

As we studied before (see (18.1)), the T-dual coordinate $\tilde{X}$ is obtained by changing the sign of the right movers in X:

$$\begin{aligned} X &= X_L(\tau + \sigma) + X_R(\tau - \sigma), \\ \tilde{X} &= X_L(\tau + \sigma) - X_R(\tau - \sigma). \end{aligned} \tag{19.26}$$

Simple relations then follow for the $\partial_\pm$ derivatives:

$$\text{Duality relations:} \quad \partial_+ X = \partial_+ \tilde{X}, \quad \partial_- X = -\partial_- \tilde{X}. \tag{19.27}$$

Since T-duality exchanges Neumann and Dirichlet boundary conditions, (19.27) can be used to obtain equations (19.23) and (19.25) from one another. Equations (19.27) hold for all values of σ, including $\sigma = 0$ and $\sigma = \pi$.

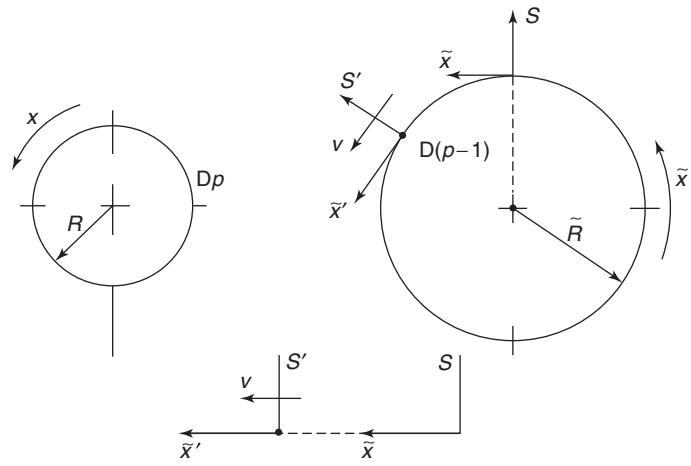

Fig. 19.1 On the left is a Dp-brane wrapped on a circle of radius R. On the right is a D$(p-1)$-brane on the dual circle of radius $\tilde{R}$. The D$(p-1)$-brane is at rest in the frame S', which is boosted with respect to the frame S.

We can now return to our present problem. We showed in Section 18.1 that a Dp-brane with one direction wrapped around a circle of radius R is T-dual to a world in which a D$(p-1)$-brane is localized at some point on a circle of radius $\tilde{R} = \alpha'/R$. Our problem is to find the dual description of a Dp-brane which, in addition to being wrapped around a compact dimension, carries an electric field that points in this direction. The boundary conditions for such a Dp-brane are given by (19.21). We claim that the dual description of this configuration is a D$(p-1)$-brane moving with constant velocity around the dual compact dimension. Our strategy will be to show that the T-dual of this moving brane configuration is described by boundary conditions that coincide with those in (19.21).

Let S be the frame which is at rest on the circle along which the D$(p-1)$-brane is moving, and let S' be the rest frame of the D$(p-1)$-brane (see Figure 19.1). The S' frame is boosted relative to the S frame by the boost parameter $\beta = v/c$, where v is the speed of the brane. In the S' frame the D$(p-1)$-brane is at rest, so we can express the boundary conditions for strings ending on the D-brane in terms of the S' string coordinates. Let X'^0 and $\tilde{X}'$ denote the string coordinates in the S' frame (primes are not σ derivatives!). Since X'^0 is Neumann and $\tilde{X}'$ is Dirichlet, we can use equation (19.25) to write

$$\partial_+ \begin{pmatrix} X'^0 \\ \tilde{X}' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \partial_- \begin{pmatrix} X'^0 \\ \tilde{X}' \end{pmatrix}. \tag{19.28}$$

We need the boundary conditions in the S frame, since this is the frame in which we know how to perform the T-duality transformation. To find them, we use the Lorentz boost

$$\begin{aligned} X'^0 &= \gamma(X^0 - \beta\tilde{X}), \\ \tilde{X}' &= \gamma(-\beta X^0 + \tilde{X}), \end{aligned} \tag{19.29}$$

where X^0 and $\tilde{X}$ are the string coordinates in the S frame and $\gamma = (1-\beta^2)^{-1/2}$. In matrix language:

$$\begin{pmatrix} X'^0 \\ \tilde{X}' \end{pmatrix} = \gamma \begin{pmatrix} 1 & -\beta \\ -\beta & 1 \end{pmatrix} \begin{pmatrix} X^0 \\ \tilde{X} \end{pmatrix} \equiv M \begin{pmatrix} X^0 \\ \tilde{X} \end{pmatrix}, \tag{19.30}$$

where we defined the constant matrix M. Substituting (19.30) into (19.28), noting that M commutes with the partial derivatives, and multiplying both sides of the equation by M^{-1}, we get

$$\partial_+ \begin{pmatrix} X^0 \\ \tilde{X} \end{pmatrix} = M^{-1} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} M \, \partial_- \begin{pmatrix} X^0 \\ \tilde{X} \end{pmatrix}. \tag{19.31}$$

This equation gives the string boundary conditions in the S frame.

Now we can perform a T-duality transformation on the $\tilde{X}$ coordinate. Using the duality relations (19.27), we see that the $\tilde{X}$ on the left-hand side can be changed into X, and the $\tilde{X}$ on the right-hand side can be changed into $(-X)$, if we include another matrix:

$$\partial_+ \begin{pmatrix} X^0 \\ X \end{pmatrix} = M^{-1} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} M \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \partial_- \begin{pmatrix} X^0 \\ X \end{pmatrix}. \tag{19.32}$$

A small calculation now gives

$$\partial_+ \begin{pmatrix} X^0 \\ X \end{pmatrix} = \begin{pmatrix} \dfrac{1+\beta^2}{1-\beta^2} & \dfrac{2\beta}{1-\beta^2} \\ \dfrac{2\beta}{1-\beta^2} & \dfrac{1+\beta^2}{1-\beta^2} \end{pmatrix} \partial_- \begin{pmatrix} X^0 \\ X \end{pmatrix}. \tag{19.33}$$

These are the open string boundary conditions for the theory dual to the moving D$(p-1)$-brane. As promised, they coincide with those in (19.21), which were written for a Dp-brane carrying an electric field, if we set

$$\mathcal{E} = 2\pi\alpha' E = \beta. \tag{19.34}$$

Our main result can thus be summarized:

> a D$(p-1)$-brane that is moving with velocity parameter β on a circle is T-dual to a Dp-brane that is wrapped on the dual circle and that carries an electric field $\mathcal{E} = \beta$ along the direction of the circle.

Furthermore, since no object with a rest mass can reach the speed of light, $\beta < 1$, and we discover that the electric field is bounded:

$$|E| = \frac{|\beta|}{2\pi\alpha'} < \frac{1}{2\pi\alpha'} = E_{\text{crit}}. \tag{19.35}$$

E_{crit} denotes the critical electric field, the maximum value of an electric field on a D-brane (in the absence of magnetic fields). Interestingly, the critical electric field coincides with the string tension:

$$E_{\text{crit}} = T_0. \tag{19.36}$$

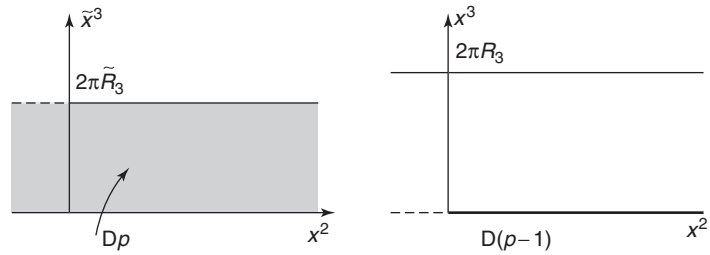

Fig. 19.2 **A Dp-brane stretched on a cylinder of circumference $2\pi\tilde{R}_3$. By performing a T-duality transformation on $\tilde{X}^3$, we obtain a D$(p-1)$-brane stretched along the axis of a dual cylinder.**

We can give an intuitive explanation for this equality by considering the motion of an open string in an electric field. As we have seen, an open string has charges of value ± 1 at the endpoints. Two forces that must cancel out act on each (zero mass) endpoint: the electric force of magnitude E and the effective tension $T_0(1 - v_\perp^2)^{1/2}$ (see (7.15)). For $E < T_0$ the two forces can be balanced for a suitable endpoint velocity $v_\perp$. When $E = T_0$ the endpoints must stop moving. Finally, for $E > T_0$ the forces at the endpoints cannot be balanced. This discussion is nicely illustrated in the detailed analysis of Problem 19.2.

19.3 D-branes with magnetic fields

We will now explore the properties of D-branes that carry magnetic fields on their world-volume. Again, our main tool will be T-duality, and we will gain considerable insight by constructing the T-dual version of a D-brane that is carrying a background magnetic field. The motion of open strings in a background magnetic field is of interest, but this subject is relegated to Problem 19.5. There you will show that an open string develops an electric dipole moment in a direction orthogonal to the motion; the magnitude of the electric dipole is proportional to both the magnetic field and the momentum of the string.

Consider a Dp-brane for which two directions of its world-volume lie on the $(x^2, \tilde{x}^3)$ plane. Assume that the dimension $\tilde{x}^3$ is compactified into a circle of radius $\tilde{R}_3$, so that x^2 and $\tilde{x}^3$ together define a cylinder of circumference $2\pi\tilde{R}_3$ (see Figure 19.2). The open string coordinates will be denoted by X^2 and $\tilde{X}^3$, and both of them are Neumann. If we perform a T-duality transformation on the string coordinate $\tilde{X}^3$, then the dual coordinate X^3 will live on a circle of radius $R_3 = \alpha'/\tilde{R}_3$. The coordinate X^3 is Dirichlet, and we now have a D$(p-1)$-brane stretching along x^2 at a fixed position $x^3 = 0$. This dual picture is shown on the right side of Figure 19.2. On the dual cylinder, the D$(p-1)$-brane appears as a line that runs parallel to the axis of the cylinder.

Now suppose that there is a nonvanishing magnetic field $F_{23} = B$ on the Dp-brane. What happens in the dual world? It turns out that the D$(p-1)$-brane tilts by an angle. Electric fields appear as boosts in the dual world; magnetic fields appear as rotations! To

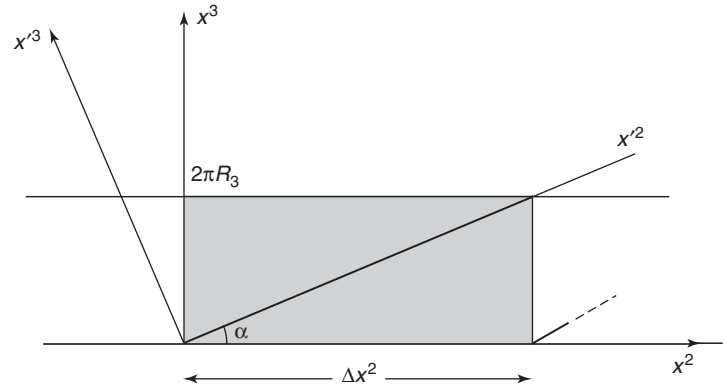

Fig. 19.3 The D-brane lies along the rotated axis x'^2. Δx^2 is the distance that one must move along the length of the cylinder for the D-brane to have wrapped once along the compact vertical direction. The rotation angle α is related to the magnetic field on the T-dual D-brane.

demonstrate this, we begin by writing the boundary conditions for open strings that end on the Dp-brane that carries the magnetic field. From (19.12) we have

$$\begin{aligned} \partial_\sigma X^2 - \mathcal{B}\partial_\tau \tilde{X}^3 &= 0, \\ \partial_\sigma \tilde{X}^3 + \mathcal{B}\partial_\tau X^2 &= 0, \end{aligned} \tag{19.37}$$

where we define the dimensionless magnetic field $\mathcal{B}$ as

$$\mathcal{B} \equiv 2\pi\alpha' B. \tag{19.38}$$

In terms of ∂_+ and ∂_- the boundary conditions take the form

$$\partial_+ \begin{pmatrix} X^2 \\ \tilde{X}^3 \end{pmatrix} = \begin{pmatrix} \dfrac{1-\mathcal{B}^2}{1+\mathcal{B}^2} & \dfrac{2\mathcal{B}}{1+\mathcal{B}^2} \\ -\dfrac{2\mathcal{B}}{1+\mathcal{B}^2} & \dfrac{1-\mathcal{B}^2}{1+\mathcal{B}^2} \end{pmatrix} \partial_- \begin{pmatrix} X^2 \\ \tilde{X}^3 \end{pmatrix}. \tag{19.39}$$

These boundary conditions encode all the effects of the magnetic field. The equations of motion for the string coordinates are the usual wave equations, which are unchanged by the presence of the magnetic field.

Quick calculation 19.4 Show that (19.39) follows from (19.3).

Consider now a D$(p-1)$-brane that is tilted on the cylinder, as illustrated in Figure 19.3. This brane wraps around the cylinder as it advances along its length. If x^3 were not compactified, then a tilted brane would be physically equivalent to a horizontal brane. But with the compactification of x^3 the tilting of the D-brane has consequences. Assume that the tilting angle is α when the magnetic field is B. Our goal is to calculate α as a function of B. The boundary conditions for the tilted D-brane are easily expressed in terms of a coordinate system x'^2, x'^3 that is rotated by the angle α. In this frame the D-brane coincides

with the x'^2 axis. Since X'^2 and X'^3 are Neumann and Dirichlet, respectively, equation (19.25) applies:

$$\partial_+ \begin{pmatrix} X'^2 \\ X'^3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \partial_- \begin{pmatrix} X'^2 \\ X'^3 \end{pmatrix}. \tag{19.40}$$

Now we perform a rotation in order to change coordinates to the unprimed frame. It is easy to figure out the proper rotation matrix from the geometry of the problem:

$$\begin{pmatrix} X'^2 \\ X'^3 \end{pmatrix} = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} X^2 \\ X^3 \end{pmatrix} \equiv R \begin{pmatrix} X^2 \\ X^3 \end{pmatrix}, \tag{19.41}$$

where we defined the rotation matrix R. Back in (19.40) we then find

$$\partial_+ \begin{pmatrix} X^2 \\ X^3 \end{pmatrix} = R^{-1} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} R \, \partial_- \begin{pmatrix} X^2 \\ X^3 \end{pmatrix}. \tag{19.42}$$

We now perform the duality transformation that takes X^3 to $\tilde{X}^3$. This is done by including an additional matrix in the right-hand side above:

$$\partial_+ \begin{pmatrix} X^2 \\ \tilde{X}^3 \end{pmatrix} = R^{-1} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} R \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \partial_- \begin{pmatrix} X^2 \\ \tilde{X}^3 \end{pmatrix}. \tag{19.43}$$

Multiplying the matrices together, we finally get

$$\partial_+ \begin{pmatrix} X^2 \\ \tilde{X}^3 \end{pmatrix} = \begin{pmatrix} \cos 2\alpha & -\sin 2\alpha \\ \sin 2\alpha & \cos 2\alpha \end{pmatrix} \partial_- \begin{pmatrix} X^2 \\ \tilde{X}^3 \end{pmatrix}. \tag{19.44}$$

Now we can compare this result with (19.39). There is a clear similarity between the two matrices: in both cases the two terms on the diagonal are the same, and the off-diagonal terms differ by only a sign. Let us solve for $\mathcal{B}$ using the diagonal terms, and then we will check that for this solution the off-diagonal terms are correct. We have

$$\frac{1-\mathcal{B}^2}{1+\mathcal{B}^2} = \cos 2\alpha, \tag{19.45}$$

which gives

$$\mathcal{B}^2 = \frac{1-\cos 2\alpha}{1+\cos 2\alpha} = \frac{1-(1-2\sin^2\alpha)}{1+(2\cos^2\alpha-1)} = \tan^2\alpha, \tag{19.46}$$

and thus we have $\mathcal{B} = \pm\tan\alpha$. We will check below that the negative sign is the correct one, so

$$\mathcal{B} = 2\pi\alpha' B = -\tan\alpha. \tag{19.47}$$

A zero magnetic field produces no rotation, and it requires an infinite magnetic field to rotate the D-brane by an angle of ninety degrees. Finally, we can use the value of $\mathcal{B}$ in (19.47) to confirm that the off-diagonal entries in the boundary conditions work out:

$$\frac{2\mathcal{B}}{1+\mathcal{B}^2} = -\frac{2\tan\alpha}{\sec^2\alpha} = -\sin 2\alpha, \tag{19.48}$$

as expected, and confirming that we chose the correct sign in (19.47). Since the boundary conditions match perfectly, we have proven that the tilted D-brane is the dual version of the D-brane with the magnetic field. Moreover, we have found the precise relation between the angle α and the magnetic field B.

To learn more from this problem, we now examine the vector potentials in the Dp-brane picture. We must find potentials A_2 and A_3 that give $F_{23} = \partial_2 A_3 - \partial_3 A_2 = B$. Let us choose

$$A_2 = 0, \quad A_3 = B\, x^2. \tag{19.49}$$

We would have found the same B had we chosen instead $A_2 = -B\, \tilde{x}^3$ and $A_3 = 0$, but we would then have had some complications because $\tilde{x}^3$ is not a well defined coordinate.

The potential A_3 lies along a compact dimension and is therefore a periodic quantity (see Section 18.3). Using (18.45) with $q = 1$, we see that

$$A_3 \sim A_3 + \frac{n}{\tilde{R}_3}, \quad n \in \mathbb{Z}. \tag{19.50}$$

This identification implies that a change in A_3 will have no physical effect if this change is quantized in units of the inverse radius $\tilde{R}_3$. It follows that the linear growth of A_3 along x^2 encodes some periodicity along the x^2 direction. The configuration must be invariant under any displacement Δx^2 that satisfies

$$\Delta A_3 = B \Delta x^2 = \frac{n}{\tilde{R}_3}. \tag{19.51}$$

The interpretation of this periodicity is quite striking in the dual world. Using dual variables, recalling (19.47), and letting $n \to -n$ for convenience, we see that

$$\Delta x^2 = -\frac{n R_3}{\alpha' B} = -\frac{2\pi n\, R_3}{2\pi \alpha' B} = \frac{2\pi n R_3}{\tan\alpha}. \tag{19.52}$$

It is convenient to rewrite this equation as

$$\tan\alpha = \frac{2\pi n R_3}{\Delta x^2}, \quad n \in \mathbb{Z}. \tag{19.53}$$

The smallest displacement Δx^2 which leads to repetition corresponds to $n = 1$:

$$\tan\alpha = \frac{2\pi R_3}{\Delta x^2}. \tag{19.54}$$

As you can see in Figure 19.3, since the D-brane is at an angle α, it completes a full wrapping of the x^3 direction precisely after moving a distance Δx^2. In this dual picture things repeat explicitly each time the D-brane completes a wrapping of the compact dimension. This provides a concrete realization of the periodicity property of the gauge potential in the original picture.

Since the relevant physics repeats along the x^2 axis, we can compactify this direction into a circle of radius R_2, where $2\pi R_2$ is the smallest repetition length:

$$2\pi R_2 = \Delta x^2 = \frac{2\pi R_3}{\tan\alpha}. \tag{19.55}$$

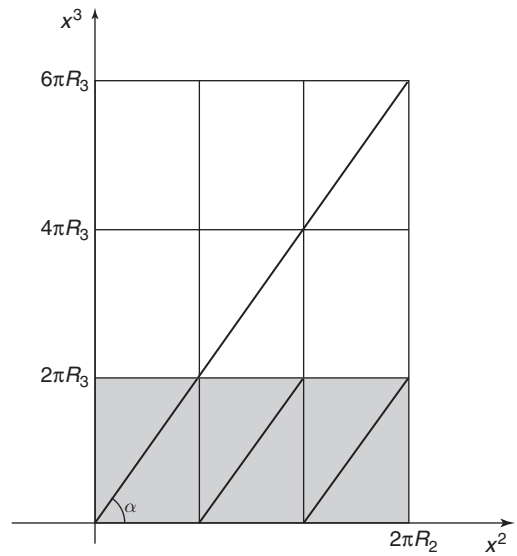

Fig. 19.4 **A D-brane wrapped around a torus of radii R_2 and R_3. The fundamental domain of the torus is shaded. The D-brane is tilted at an angle α such that, as it goes around the x^2 circle, it wraps three times around the x^3 circle.**

This turns the cylinder around which the D-brane wraps into a torus with radii R_2 and R_3. The D-brane, once infinitely long, is now of finite length. It wraps diagonally across the fundamental domain of the torus, which is shown as shaded in Figure 19.3. Equivalently, as it wraps once around x^2, it also wraps once around x^3. The Dp-brane world also has a torus, with radii R_2 – since we do not T-dualize in this direction – and $\tilde{R}_3$.

Suppose now that on the torus where the D-brane wraps diagonally we change the angle α in such a way that

$$\tan\alpha = n\,\frac{2\pi\,R_3}{2\pi\,R_2} = n\,\frac{R_3}{R_2}, \tag{19.56}$$

with $n > 1$. In Figure 19.4 we show the case $n = 3$. Equation (19.56) represents the situation in which the D-brane wraps n times around the x^3 direction as it wraps once around the x^2 direction. What is the significance of the integer n in the dual world with the magnetic field? To find out, we calculate the total magnetic flux Φ on the torus. The flux is simply the magnetic field multiplied by the area of the fundamental domain:

$$\Phi = B\,(2\pi\,\tilde{R}_3)\,(2\pi\,R_2). \tag{19.57}$$

Using the value of the magnetic field (19.47), writing $\tilde{R}_3$ in terms of R_3, and using (19.56), we can simplify the expression for the flux:

$$\Phi = -\frac{\tan\alpha}{2\pi\alpha'}\,\frac{2\pi\alpha'}{R_3}\,2\pi\,R_2 = -2\pi\,\frac{R_2}{R_3}\,\tan\alpha = -2\pi n. \tag{19.58}$$

This means that the magnetic flux is quantized! Since B is uniform, it must be quantized as well. Geometrically, this quantization emerges because, on a given torus, a D-brane

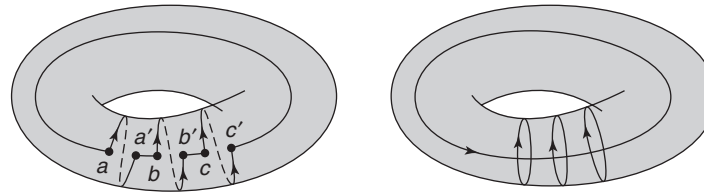

Fig. 19.5 A D-brane wrapping three times around the short circle of a torus as it travels along the long circle of the torus. In the limit that $a \to a'$, $b \to b'$, and $c \to c'$, the figure becomes one in which there are three D-branes wrapped around the short circle and one D-brane wrapped around the long circle.

can only be tilted at specific angles if it is to close after one winding in the x^2 direction. Actually, one can derive the quantization of the magnetic flux on a torus directly (see Problem 19.3), and there is no need to assume that the magnetic field is constant over the torus. What happens in our context if the magnetic field is not constant? A nonconstant, but static, magnetic field does not satisfy the relevant equations of electromagnetism, so this makes the question about the T-dual configuration somewhat ambiguous. The T-dual D-brane on the dual torus is not expected to appear as a straight line but rather as a curve. Because of the brane tension, this configuration cannot be static either.

There is a sense in which a tilted D$(p-1)$-brane that wraps n times around x^3 as it wraps once around x^2 is in the same class as a configuration composed of one D$(p-1)$-brane that only wraps in the x^2 direction and n D$(p-1)$-branes that only wrap in the x^3 direction:

$$\text{D}(p-1)\text{-brane (along } x^2) \quad \text{and} \quad n\,\text{D}(p-1)\text{-branes (along } x^3). \tag{19.59}$$

Consider Figure 19.5, which shows the torus when $n = 3$. We have deformed the D-brane so that it is no longer a straight line. Most of the D-brane trajectory is horizontal except for the wrappings around x^3, which now occur rapidly. Since we can achieve this configuration by a continuous deformation of the original straight D-brane, this must preserve the quantized flux on the dual brane, although the B field would not be expected to remain constant. In the limit that $a \to a'$, $b \to b'$, and $c \to c'$, the D-brane turns into a single horizontal D-brane and three vertical D-branes. In general, we have the set of branes indicated in (19.59). The configuration in (19.59) may be static since the branes are once again straight lines on the torus. The original tilted D-brane configuration and the final configuration (19.59) are related by a deformation process. They are not physically equivalent configurations; instead, they are deformation equivalent, meaning that they can be deformed into each other. The precise mathematical statement is that the curves defining the two brane configurations are homologically equivalent.

Now consider performing, once again, a T-duality transformation in the x^3 direction. Previously, T-duality turned the tilted D$(p-1)$-brane into a Dp-brane with a magnetic field. This time the result of T-duality on (19.59) is different. The D$(p-1)$-brane along x^2 becomes a Dp-brane in the dual world. The n D$(p-1)$-branes in the x^3 direction,

however, become n D$(p-2)$-branes, located at certain fixed points on the torus. So in the dual world we have one Dp-brane and n D$(p-2)$-branes:

$$\text{D}p\text{-brane} \quad \text{and} \quad n\,\text{D}(p-2)\text{-branes}. \tag{19.60}$$

This shows that the original Dp-brane with a constant magnetic field is deformation equivalent to a Dp-brane with a number n of D$(p-2)$-branes on its world-volume. A D2-brane with a constant magnetic field of flux $2\pi n$, for example, is deformation equivalent to a D2-brane with n D0-branes on its world-volume. The deformation that takes the latter into the former is the process where the D0-branes *dissolve* into the torus. The physical picture is clear: D0-branes on a D2-brane represent a magnetic field that vanishes everywhere except at the positions of the D0-branes, where it has infinite magnitude but finite flux. A constant magnetic field with the same flux is possible, and the spread-out magnetic field represents the dissolved D0-branes. The technical implication of the word dissolved is that the resulting configuration is a bound state: it has less energy than the total energy of the constituent D2- and D0-branes taken separately. The energies actually add in quadrature, as you will prove in Problem 19.4. An extremely strong magnetic field can be viewed as a D2-brane with infinitely many dissolved D0-branes. This was the picture suggested by a direct analysis of the boundary conditions at the end of Section 19.1.

Problems

Problem 19.1 Constant electromagnetic fields on D-branes from constant $B_{\mu\nu}$ in spacetime.

We would like to show that a Dp-brane with a constant electromagnetic background field $\bar{F}_{mn}$ is equivalent to a Dp-brane with $F_{mn} = 0$ but in a spacetime with a constant Kalb–Ramond field $B_{mn} = \bar{F}_{mn}$. The indices m, n are brane indices.

(a) Refer to Section 16.3, and consider a situation where $B_{mn} = \bar{F}_{mn}$ is a nonzero constant and $F_{mn} = 0$. Find an explicit gauge parameter for which the transformed fields are $B_{mn} = 0$ and $F_{mn} = \bar{F}_{mn}$. Note that, as expected, the gauge invariant field strength $\mathcal{F}_{mn}$ is unchanged.

(b) Let B_{mn} be a constant Kalb–Ramond field. Consider the action (16.45), and show that the integrand is a total derivative. Drop the total τ derivatives, and show that for open strings the total σ derivatives contribute

$$S_B = \frac{1}{2}\int d\tau\, B_{mn}\Big(X^m \partial_\tau X^n|_{\sigma=\pi} - X^m \partial_\tau X^n|_{\sigma=0}\Big).$$

Compare with (19.3) and comment.

Problem 19.2 Motion of an open string in a constant electric field.

Consider the motion of an open string in the background of an electric field of magnitude $E = \mathcal{E}/2\pi\alpha'$ that points in the x^1 direction. When $\mathcal{E} \to 0$ the motion is rigid rotation in the (x^2, x^3) plane. We use string coordinates $X^\mu = (X^0, X^1, \vec{X})$ where $\vec{X} = (X^2, X^3)$.

All other coordinates are set equal to zero. The classical solution must satisfy the boundary conditions

$$\partial_\sigma X^0 - \mathcal{E}\,\partial_\tau X^1 = 0, \quad \partial_\sigma X^1 - \mathcal{E}\,\partial_\tau X^0 = 0, \quad \partial_\sigma \vec{X} = 0, \tag{1}$$

as well as the wave equations and constraints:

$$(\partial_\tau^2 - \partial_\sigma^2) X^\mu = 0, \quad \dot{X}\cdot X' = 0, \quad \dot{X}^2 + X'^2 = 0. \tag{2}$$

We look for a solution in which the string rotates around the x^1 axis with angular frequency ω but is not restricted to lie on the (x^2, x^3) plane. Construct the solution beginning with the ansatz

$$(X^0, X^1, X^2, X^3) = \Big(\alpha\,\tau,\ \beta(\sigma - \tfrac{\pi}{2}), \gamma\cos\sigma\cos\tau,\ \ \gamma\cos\sigma\sin\tau\Big). \tag{3}$$

(a) Examine equations (1) and (2) and fix the constants α, β, and γ in terms of ω and $\mathcal{E}$.

(b) Give $\vec{X}(t,\sigma)$, and show that the velocity of the endpoints is $v = \sqrt{1-\mathcal{E}^2}$. Show also that the string is stretched along the x^1 axis a total distance $\Delta X^1 = \pi\mathcal{E}/\omega$.

(c) Plot the string in the (x^1, x^2) plane at $t = 0$. Show that at the endpoints, the string points in the direction of the electric field. Interpret this result by verifying that the effective string tension cancels the electric force at the endpoints. Show that at the origin

$$\frac{dX^2}{dX^1}\Big|_{X^1=0} = -\sqrt{\frac{1}{\mathcal{E}^2} - 1}. \tag{4}$$

Does this make sense for $0 \le \mathcal{E} \le 1$? Describe the solution in the limit as $\mathcal{E} \to 1$ with ω finite.

(d) Prove that the string energy $U = \int T_0 ds (1 - v_\perp^2)^{-1/2}$ is equal to $\pi T_0/\omega$, for all values of the electric field.

The front cover of the book shows three surfaces generated by the motion of open strings. The surfaces are obtained for $\mathcal{E} = 0.3$, 0.6, and 0.95. Construct the associated plot of the three strings at $t = 0$ on the (x^1, x^2) plane, setting $\omega = 1$. The back cover shows the strings on the (x^1, x^2) plane at $t = 0$ and half a period later. The front cover shows the surfaces generated by the motion of the strings.

Problem 19.3 Quantization of magnetic flux on two-tori.

In this problem you will show that the flux Φ of the magnetic field on a two-torus is quantized: $\Phi = 2\pi n$, with $n \in \mathbb{Z}$. This quantization emerges when we try to construct a consistent vector potential. We assume, of course, that the Maxwell gauge group is $U(1)$ and, as explained in Section 18.2, the gauge parameter is $U = \exp(i\chi)$ (with $q = c = \hbar = 1$).

Consider a two-torus of length L_x along the x axis and length L_y along the y axis. Let $F_{12} = \partial_x A_y - \partial_y A_x = B$, where B is the value of the magnetic field.

(a) Assume the magnetic field $B = B_0$ is constant. Take $A_y(x,y) = B_0 x$ and $A_x = 0$. Note that $A_y(x + L_x, y) \neq A_y(x,y)$. Allow for $A_y(x + L_x, y)$ to be gauge equivalent to $A_y(x,y)$. Show that the condition of having a well defined gauge parameter leads to the desired quantization.

(b) Consider the flux integral of the magnetic field (not necessarily a constant one), and use Stokes' theorem to relate it to a line integral of the vector potential around the four sides of the rectangle $0 \le x \le L_x$, $0 \le y \le L_y$. Use the gauge transformation properties of gauge-field line integrals to argue that the magnetic flux is properly quantized.

Problem 19.4 Dissolved D$(p-2)$-branes.

Consider a Dp-brane that is wrapped on a torus with radii R_2 and $\tilde{R}_3$ and carries a magnetic field with flux $|\Phi| = 2\pi n$ (as in Section 19.3). The other $p-2$ dimensions on the brane are wrapped around a compact space of volume V_{p-2}. Assume that the string coupling constant takes the value g. The energy E of this configuration is equal to the energy of the T-dual configuration (dualized along $\tilde{R}_3$) where a D$(p-1)$ brane wraps along the dual torus. Show that

$$E = T_{p-1}(\tilde{g})\ V_{p-2}\sqrt{(2\pi R_2)^2 + (2\pi R_3 n)^2}, \tag{1}$$

where $\tilde{g}$ is the string coupling in the dual configuration. Now rewrite E in terms of variables in the original picture with the Dp-brane and the magnetic field, showing that

$$E = \sqrt{M_p^2 + (nM_{p-2})^2}, \tag{2}$$

where M_p is the mass of the Dp-brane and M_{p-2} is the mass of a D$(p-2)$ brane. Explain why this result shows that the energy E is smaller than the sum of the energies of the Dp-brane and n D$(p-2)$-branes. This justifies the description of the configuration as one where n D$(p-2)$-branes have *dissolved* inside the Dp-brane.

Problem 19.5 Motion of an open string in a constant magnetic field.

Consider the following expansion for a spatial open string coordinate on the world-volume of a D-brane with a constant magnetic field:

$$X_i(\tau,\sigma) = x_i + 2\alpha'(p_i\tau + 2\pi\alpha' F_{ij}\, p^j\sigma). \tag{1}$$

The coordinates X_i clearly solve the wave equations. Moreover, assume that

$$X^0 = 2\alpha' p^0\tau. \tag{2}$$

(a) Prove that, for a calculable value of p^0 that you should determine, the ansatz in (1) and (2) satisfies: the boundary conditions (19.11), the constraint $\dot{X}\cdot X' = 0$, and the constraint $\dot{X}^2 + X'^2 = 0$.

(b) Let $\Delta X_i = X_i(\tau,\pi) - X_i(\tau,0)$ denote the "span" of the string. Verify that

$$\Delta X_i = (2\pi\alpha')^2 F_{ij}\, p^j \quad \text{and} \quad p^i\,\Delta X_i = 0.$$

Since the electric dipole of an open string is a vector parallel to its span (why?), we conclude that the dipole is perpendicular to the momentum.

(c) For motion in three space dimensions, we have $F_{ij} = \epsilon_{ijk}B_k$. Show that

$$\Delta\vec{X} = (2\pi\alpha')^2\ \vec{p}\times\vec{B}.$$

The dipole moment is orthogonal to both $\vec{B}$ and $\vec{p}$.

(d) Let $\mathcal{B}_k = 2\pi\alpha' B_k$. Show that the velocity of the string is given by

$$\vec{v} = \frac{\vec{n}}{\sqrt{1 + \vec{\mathcal{B}} \cdot \vec{\mathcal{B}} - (\vec{\mathcal{B}} \cdot \vec{n})^2}},$$

where $\vec{n}$ is a unit vector in the direction of the momentum. Note that when $B_k \to 0$ we have $|\vec{v}| \to 1$, so that the string approaches the speed of light.

(e) In flat space the dot product of two vectors $\vec{a}$ and $\vec{b}$ is defined by $\vec{a} \cdot \vec{b} = a^i b^j \eta_{ij} = a^i b^i$. Now introduce a new product $*$ defined by

$$\vec{a} * \vec{b} = a^i b^j \, \bar{\eta}_{ij}, \text{ where } \bar{\eta}_{ij} \equiv \eta_{ij}(1 + \vec{\mathcal{B}} \cdot \vec{\mathcal{B}}) - \mathcal{B}_i \mathcal{B}_j.$$

Show that the mass-shell condition for the string can be written as

$$-(p^0)^2 + \vec{p} * \vec{p} = 0.$$

Moreover, show that the velocity of the string satisfies

$$\vec{v} * \vec{v} = 1.$$

The metric $\bar{\eta}_{ij}$ is called the open string metric. It is a natural metric for the physics of open strings in the presence of a magnetic field. In the open string metric, the motion studied here shares the properties of free open string motion.

20 Nonlinear and Born–Infeld electrodynamics

We introduce nonlinear theories of electrodynamics, which generalize the linear Maxwell theory. Born–Infeld electrodynamics is a specific theory of nonlinear electrodynamics with particularly nice properties. It incorporates maximal electric fields, and in it point charges have finite electrostatic self-energy. We use T-duality arguments to explain why electromagnetic fields on the world-volumes of D-branes are governed by Born–Infeld theory.

20.1 The framework of nonlinear electrodynamics

Maxwell's equations are both the basis for classical electromagnetism and the starting point for the formulation of quantum electrodynamics, a theory that has been tested to a high degree of accuracy. Maxwell's equations are written in terms of electric and magnetic fields, which in turn arise from gauge potentials. Charges and currents are the sources in Maxwell's equations.

A somewhat different version of Maxwell's equations is used in the study of electromagnetic phenomena in the presence of materials. In this case, the materials contribute polarization charges to the charge density and magnetization currents to the current density. The original Maxwell equations hold, but one must include these contributions to the charges and currents. This is done efficiently by introducing, in addition to $\vec{E}$ and $\vec{B}$, the fields $\vec{D}$ and $\vec{H}$. The equations of electromagnetism in the presence of materials then take the form

$$\begin{aligned} \nabla \times \vec{E} &= -\frac{1}{c}\frac{\partial \vec{B}}{\partial t}\,, \\ \nabla \cdot \vec{B} &= 0\,, \end{aligned} \tag{20.1}$$

for the equations without sources, and

$$\begin{aligned} \nabla \cdot \vec{D} &= \rho\,, \\ \nabla \times \vec{H} &= \frac{\vec{j}}{c} + \frac{1}{c}\frac{\partial \vec{D}}{\partial t}\,, \end{aligned} \tag{20.2}$$

for the equations with sources. Here ρ and $\vec{j}$ are called *free* sources. This means that they *do not* take into account polarization charges or magnetization currents. The free

sources are charges and currents that are not bound to the materials. The contributions of polarization and magnetization have been incorporated through the fields $\vec{D}$ and $\vec{H}$. In fact, given a material, there are phenomenological relations that express $\vec{D}$ and $\vec{H}$ in terms of $\vec{E}$ and $\vec{B}$:

$$\vec{D} = \vec{D}(\vec{E}, \vec{B}) , \quad \vec{H} = \vec{H}(\vec{E}, \vec{B}) . \tag{20.3}$$

Without these relations the two sets of equations in (20.1) and (20.2) would be unrelated. Linear dielectrics, for example, obey $\vec{D} = \epsilon \vec{E}$ (and $\vec{H} = \vec{B}$), where the constant ϵ is the electric permittivity. Linear magnetic materials obey $\vec{B} = \mu \vec{H}$ (and $\vec{D} = \vec{E}$), where the constant μ is the magnetic permeability. For more complicated materials the relations need not be linear. In such cases, equations (20.1), (20.2), and (20.3) define a nonlinear theory of electrodynamics.

In introductory courses on electromagnetism the point is often made that the "fundamental" Maxwell equations are those in terms of $\vec{E}$ and $\vec{B}$, together with the full charge and current densities. The equations above are regarded to be of limited validity, as they deal with materials, most of which do not generally lend themselves to exact analysis. Born–Infeld and related nonlinear theories of electrodynamics in fact suggest that the above equations are as fundamental as the original Maxwell equations, if not more so. These theories do not aim to describe electromagnetism in the presence of materials but rather electromagnetism in the vacuum. The point is that in nonlinear electrodynamics the vacuum itself behaves as some kind of material. As we will show, general Lagrangians lead directly to equations (20.2), together with nontrivial relationships between $(\vec{D}, \vec{H})$ and $(\vec{E}, \vec{B})$.

In nonlinear electrodynamics the electromagnetic fields $\vec{E}$ and $\vec{B}$ are still encoded in the field strength $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$, as shown in equation (3.20). Using spatial indices i, j, we write

$$F_{i0} = E_i , \quad F_{ij} = \epsilon_{ijk} B_k . \tag{20.4}$$

Equations (20.1) are automatically satisfied because $\vec{E}$ and $\vec{B}$ arise from potentials. How do we write equations (20.2)? Recall that the similar looking equations in Maxwell theory were written in (3.34): $\partial_\nu F^{\mu\nu} = (1/c) j^\mu$. It follows that (20.2) are given by

$$\frac{\partial G^{\mu\nu}}{\partial x^\nu} = \frac{1}{c} j^\mu , \tag{20.5}$$

where $G^{\mu\nu} = -G^{\nu\mu}$ is obtained from $F^{\mu\nu}$ by replacing $\vec{E}$ by $\vec{D}$ and $\vec{B}$ by $\vec{H}$:

$$G^{\mu\nu} = \begin{pmatrix} 0 & D_x & D_y & D_z \\ -D_x & 0 & H_z & -H_y \\ -D_y & -H_z & 0 & H_x \\ -D_z & H_y & -H_x & 0 \end{pmatrix} . \tag{20.6}$$

Although the matrix $G^{\mu\nu}$ written here applies only in four-dimensional spacetime, equations (20.5) make sense in any number of dimensions. Together with the definition of $F_{\mu\nu}$

in terms of potentials, they form the equations of nonlinear electrodynamics in an arbitrary number of dimensions. Of course, the relation between $G^{\mu\nu}$ and $F^{\mu\nu}$ must also be given.

We now show that equation (20.5) follows from the variation of an action. Moreover, the definition of $G^{\mu\nu}$ in terms of the field strength $F^{\mu\nu}$ also arises. Consider the action S, written for arbitrary spacetime dimensionality:

$$S = \int d^D x \, \mathcal{L}\,(F_{\mu\nu}) + \frac{1}{c}\int d^D x \, A_\mu \, j^\mu \,. \tag{20.7}$$

For simplicity, the Lagrangian density $\mathcal{L}(F_{\mu\nu})$ is assumed to depend only on the field strength and not, for example, on its derivatives. Since the field strength is gauge invariant, $\mathcal{L}$ is also gauge invariant. The Lagrangian density $\mathcal{L}$ is otherwise arbitrary; it need not coincide with the Maxwell Lagrangian density. In order to perform the variation of the action efficiently, we need to *define* partial derivatives with respect to field strengths.

For this purpose we first note that the variations $\delta F_{\mu\nu}$ are constrained by antisymmetry: $\delta F_{\mu\nu} = -\delta F_{\nu\mu}$. For any function M of the field strengths we then write

$$\delta M = \frac{1}{2}\,\frac{\partial M}{\partial F_{\mu\nu}}\,\delta F_{\mu\nu} \,. \tag{20.8}$$

As usual, repeated indices are summed over. Since the variations $\delta F_{\mu\nu}$ are antisymmetric, we can require that

$$\frac{\partial M}{\partial F_{\mu\nu}} = -\frac{\partial M}{\partial F_{\nu\mu}} \,. \tag{20.9}$$

Equations (20.8) and (20.9) together define the partial derivatives $\partial M/\partial F_{\mu\nu}$.

Quick calculation 20.1 Use $M = F_{12}$ and the above equations to prove that $\partial F_{12}/\partial F_{12} = 1$ and that $\partial F_{12}/\partial F_{\mu\nu} = 0$ if $(\mu, \nu) \neq (1, 2)$ and $(\mu, \nu) \neq (2, 1)$.

It is also useful to learn how to use the chain rule. In order to calculate derivatives of M with respect to some variable U we write

$$\delta M = \frac{1}{2}\,\frac{\partial M}{\partial F_{\mu\nu}}\,\frac{\partial F_{\mu\nu}}{\partial U}\,\delta U \equiv \frac{\partial M}{\partial U}\delta U \,, \tag{20.10}$$

thus learning that

$$\frac{\partial M}{\partial U} = \frac{1}{2}\,\frac{\partial M}{\partial F_{\mu\nu}}\,\frac{\partial F_{\mu\nu}}{\partial U} \,. \tag{20.11}$$

We are now in a position to vary the action S. Using (20.8) with M set equal to $\mathcal{L}$, the variation of the first term in S gives

$$\delta \int d^D x \, \mathcal{L} = \int d^D x \, \frac{1}{2}\,\frac{\partial \mathcal{L}}{\partial F_{\mu\nu}}\delta F_{\mu\nu} = \int d^D x \, \frac{1}{2}\,\frac{\partial \mathcal{L}}{\partial F_{\mu\nu}}(\partial_\mu \delta A_\nu - \partial_\nu \delta A_\mu) \,. \tag{20.12}$$

Integrating by parts, relabeling indices, and using (20.9) we find

$$\delta \int d^D x \, \mathcal{L} = \int d^D x \, \delta A_\mu \, \partial_\nu \Big(\frac{\partial \mathcal{L}}{\partial F_{\mu\nu}} \Big) \,. \tag{20.13}$$

The variation of the whole action (20.7) then gives

$$\delta S = \int d^D x \, \delta A_\mu \Big[\partial_\nu \Big(\frac{\partial \mathcal{L}}{\partial F_{\mu\nu}} \Big) + \frac{1}{c} \, j^\mu \Big] . \tag{20.14}$$

Equating this variation to zero, we find (20.5) if we identify

$$G^{\mu\nu} \equiv - \frac{\partial \mathcal{L}}{\partial F_{\mu\nu}} . \tag{20.15}$$

The tensor $G^{\mu\nu}$ is antisymmetric by construction. If the nonlinear Lagrangian is known, (20.15) expresses $G^{\mu\nu}$ as a function of the field strength $F_{\mu\nu}$. To illustrate this, let us calculate $\vec{D}$. Equation (20.6) tells us that $G^{0i} = D_i$, and, for arbitrary spacetime dimensionality, this is taken to be the definition of the vector $\vec{D}$. We begin by calculating the derivative $\partial \mathcal{L} / \partial E_i$. Using the chain rule (20.11) and recalling that $E_i = F_{i0} = -F_{0i}$, we find

$$\frac{\partial \mathcal{L}}{\partial E_i} = \frac{1}{2} \frac{\partial \mathcal{L}}{\partial F_{0i}} \frac{\partial F_{0i}}{\partial E_i} + \frac{1}{2} \frac{\partial \mathcal{L}}{\partial F_{i0}} \frac{\partial F_{i0}}{\partial E_i} = - \frac{\partial \mathcal{L}}{\partial F_{0i}} = G^{0i} = D_i \, . \tag{20.16}$$

Therefore, we have shown that

$$\vec{D} = \frac{\partial \mathcal{L}}{\partial \vec{E}} . \tag{20.17}$$

In four-dimensional spacetime we can also write $\vec{H}$ in terms of derivatives of $\mathcal{L}$. For example, since $B_1 = F_{23} = -F_{32}$, we have

$$\frac{\partial \mathcal{L}}{\partial B_1} = - \frac{1}{2} \frac{\partial \mathcal{L}}{\partial F_{32}} + \frac{1}{2} \frac{\partial \mathcal{L}}{\partial F_{23}} = G^{32} = -H_x \, . \tag{20.18}$$

Working out the other two components, we find that the end result is

$$\vec{H} = - \frac{\partial \mathcal{L}}{\partial \vec{B}} . \tag{20.19}$$

If $\mathcal{L}(\vec{E}, \vec{B})$ is known, equations (20.17) and (20.19) give us $\vec{D}$ and $\vec{H}$. For Maxwell electrodynamics, for example, the Lagrangian density is

$$\mathcal{L}_M = - \frac{1}{4} F^{\mu\nu} F_{\mu\nu} = \frac{1}{2} \, (E^2 - B^2) \, . \tag{20.20}$$

Here we defined $E^2 = \vec{E} \cdot \vec{E}$ and $E = |\vec{E}|$, with similar definitions for $\vec{B}$, $\vec{H}$, and $\vec{D}$. It follows from (20.20) that $\vec{D} = \vec{E}$ and $\vec{H} = \vec{B}$.

$\vec{E}$ is related to the time derivative of $\vec{A}$, so it may be viewed as a velocity. Equation (20.17) then implies that $\vec{D}$ is the canonical momentum associated with the velocity $\vec{E}$. This suggests that the Hamiltonian, or energy functional, is given as

$$\mathcal{H} = \vec{D} \cdot \vec{E} - \mathcal{L} \,. \tag{20.21}$$

This formula is clearly correct for Maxwell theory: since $\vec{D} = \vec{E}$, we find $\mathcal{H} = \frac{1}{2}(E^2 + B^2)$, which is the familiar energy density of the electromagnetic field. In Problem 20.1 you will give a proof that $\mathcal{H}$ is indeed a conserved energy for arbitrary $\mathcal{L}$. While some of our results in this section are written in the language of four dimensions, the important ideas work in all dimensions. In fact, all of the shaded equations are valid in any spacetime dimension.

Let us conclude this section by considering some candidate Lagrangian densities in four-dimensional spacetime. The density $\mathcal{L}$ in (20.7) must be both gauge invariant and Lorentz invariant. Since it is built from field strengths it is clearly gauge invariant. To be Lorentz invariant the field strengths must form objects with no free indices. There are two independent nontrivial Lorentz invariant objects that we can build using $F_{\mu\nu}$ and none of its derivatives:

$$\begin{aligned} s &\equiv -\frac{1}{4} F^{\mu\nu} F_{\mu\nu} = \frac{1}{2}(E^2 - B^2), \\ p &\equiv -\frac{1}{4} \tilde{F}^{\mu\nu} F_{\mu\nu} = \vec{E} \cdot \vec{B}. \end{aligned} \tag{20.22}$$

In fact, it can be proven that s and p are the *only* independent invariants that can be built from $F_{\mu\nu}$. The invariant p makes use of the *dual* field strength $\tilde{F}^{\mu\nu}$, which is defined as

$$\tilde{F}^{\mu\nu} \equiv \frac{1}{2} \epsilon^{\mu\nu\rho\sigma} F_{\rho\sigma} \,. \tag{20.23}$$

Here $\epsilon^{\mu\nu\rho\sigma}$ is a totally antisymmetric Lorentz tensor (more precisely, it is a pseudo-tensor). This implies that $\tilde{F}^{\mu\nu}$ is antisymmetric. Just like $\eta^{\mu\nu}$, the tensor $\epsilon^{\mu\nu\rho\sigma}$ takes the same values in all Lorentz frames. In any frame $\epsilon^{0123} = 1$, and $\epsilon^{\mu\nu\rho\sigma}$ vanishes if any index is repeated. For example,

$$\tilde{F}^{01} = \frac{1}{2} \epsilon^{01\rho\sigma} F_{\rho\sigma} = \frac{1}{2}(F_{23} - F_{32}) = F_{23} = B_x \,. \tag{20.24}$$

Calculating all the other entries, we have

$$\tilde{F}^{\mu\nu} = \begin{pmatrix} 0 & B_x & B_y & B_z \\ -B_x & 0 & -E_z & E_y \\ -B_y & E_z & 0 & -E_x \\ -B_z & -E_y & E_x & 0 \end{pmatrix} . \tag{20.25}$$

Quick calculation 20.2 Calculate the other entries in the matrix $\tilde{F}^{\mu\nu}$.

Quick calculation 20.3 Show that p takes the value quoted in (20.22).

In four dimensions, the most general Lorentz invariant Lagrangian density built out of $F_{\mu\nu}$ is an arbitrary function of s and p. The Maxwell Lagrangian density is simply s.

20.2 Born–Infeld electrodynamics

In this section we will write the Lagrangian that turns out to describe the electromagnetic fields that live on the world-volumes of D-branes. This is the Born–Infeld electromagnetic Lagrangian. In Born–Infeld theory, as we shall see, the electrostatic self-energy of a point charge is finite. This is an improvement over Maxwell theory, where the self-energy of a point charge is infinite. The endpoints of open strings are point charges, so it is reassuring to see that in string theory these do not carry infinite energy. In the following section we will explain how T-duality gives direct evidence that the Born–Infeld Lagrangian governs the dynamics of electromagnetic fields on D-branes.

We begin our work in four-dimensional spacetime. In addition to the natural requirements of gauge and Lorentz invariance, we impose two constraints on the nonlinear Lagrangian. First, it must reduce to the Maxwell Lagrangian for small $\vec{E}$ and $\vec{B}$. Second, there should be a maximal electric field when $\vec{B} = 0$. We showed in Section 19.2 that string theory has a critical electric field $E_{\text{crit}} = 1/(2\pi\alpha') \equiv b$. For the originators of nonlinear electrodynamics, a maximal electric field was needed to obtain point charges with finite self-energy.

How can we impose the existence of a maximal electric field on our Lagrangian? In special relativity the maximal velocity is clearly evident in the point particle Lagrangian (5.8): the argument of the square root must be positive, and this ensures that the particle velocity cannot exceed the velocity of light. To impose $E \leq b$ we also use a square root. There is a simple expression which uses only the Lorentz invariant s:

$$\mathcal{L} = -b^2\sqrt{1 - \frac{(E^2 - B^2)}{b^2}} + b^2 = -b^2\sqrt{1 - \frac{2s}{b^2}} + b^2 \,. \tag{20.26}$$

As required, $E \leq b$ for $B = 0$. Moreover, for small fields we have $s \ll b^2$, so

$$\mathcal{L} = -b^2\Big(1 - \frac{s}{b^2}\Big) + b^2 + \mathcal{O}(s^2) = s + \mathcal{O}(s^2) \,. \tag{20.27}$$

For small fields we recover the Maxwell Lagrangian. While (20.26) satisfies the two conditions we requested, a somewhat more complicated Lagrangian has even nicer features. Consider the Born–Infeld Lagrangian density

$$\mathcal{L} = -b^2\sqrt{1 - \frac{E^2 - B^2}{b^2} - \frac{(\vec{E}\cdot\vec{B})^2}{b^4}} + b^2 = -b^2\sqrt{1 - \frac{2s}{b^2} - \frac{p^2}{b^4}} + b^2 \,. \tag{20.28}$$

Since it is built from s and p, this density is also Lorentz invariant. For small fields where s and p are comparable and both are much smaller than b^2, the weak field approximation is not changed: $\mathcal{L} \sim s$. In a generic theory of nonlinear electrodynamics, waves with different polarizations propagate with different velocities through a background electromagnetic field. In Born–Infeld theory the velocity is independent of the polarization. In all theories of nonlinear electrodynamics, Born–Infeld included, there are nontrivial dispersion relations; that is, waves of different frequencies travel with different velocities.

The Born–Infeld Lagrangian density is special for yet another reason. It can be written elegantly in terms of the square root of a determinant:

$$\mathcal{L} = -b^2\sqrt{-\det\Big(\eta_{\mu\nu} + \frac{1}{b}F_{\mu\nu}\Big)} + b^2\,. \tag{20.29}$$

As opposed to (20.28), this formula allows an obvious generalization to any number of dimensions. A calculation is needed, however, to show that the two densities are the same in four dimensions. The calculation is straightforward but a little long: one must simply write the 4-by-4 matrix and calculate its determinant. It is more instructive to understand why the equality is reasonable. To simplify the equations, we temporarily set $b = 1$ in (20.29):

$$\mathcal{L} = -\sqrt{-\det\big(\eta_{\mu\nu} + F_{\mu\nu}\big)} + 1\,. \tag{20.30}$$

The original Lagrangian density is recovered by letting $F_{\mu\nu} \to F_{\mu\nu}/b$ and $\mathcal{L} \to b^2\mathcal{L}$. The Lagrangian (20.28) is explicitly Lorentz invariant. How do we see the Lorentz invariance of (20.30)? To show the Lorentz invariance of $\det(\eta_{\mu\nu} + F_{\mu\nu})$ we first note that the determinant of an arbitrary matrix M with components $M_{\mu\nu}$ is the same as the determinant of the matrix $\bar{M}$ with components $M^{\mu\nu}$. To see this, we first write

$$M_{\mu\nu} = \eta_{\mu\rho}\eta_{\nu\sigma}M^{\rho\sigma} = \eta_{\mu\rho}M^{\rho\sigma}\eta_{\sigma\nu} \;\longrightarrow\; M = \eta\bar{M}\eta\,, \tag{20.31}$$

where η is the matrix with components $\eta_{\mu\nu}$. Taking determinants of both sides of the equation, we confirm that

$$\det M = \det(\eta\bar{M}\eta) = (\det\bar{M})(\det\eta)^2 = \det\bar{M}\,. \tag{20.32}$$

Let $\bar{\eta}$, F, and $\bar{F}$ denote the matrices with entries $\eta^{\mu\nu}$, $F_{\mu\nu}$, and $F^{\mu\nu}$, respectively. On account of the result just established, we have

$$\det(\eta + F) = \det(\bar{\eta} + \bar{F})\,. \tag{20.33}$$

It therefore suffices to prove the Lorentz invariance of $\det(\bar{\eta} + \bar{F})$.

Consider a Lorentz transformation $x'^{\mu} = L^{\mu}_{\ \nu}x^{\nu}$, as written in (2.38). The matrix L, with entries $L^{\mu}_{\ \nu}$ (μ is the row index and ν is the column index), satisfies $(\det L)^2 = 1$. Since both $\eta^{\mu\nu}$ and $F^{\mu\nu}$ are Lorentz tensors of the same type, they transform under Lorentz transformations in the same way. The tensors carry two indices, so the matrix L must be applied to them twice:

$$\eta'^{\mu\nu} + F'^{\mu\nu} = L^{\mu}_{\ \rho}L^{\nu}_{\ \sigma}(\eta^{\rho\sigma} + F^{\rho\sigma}) = L^{\mu}_{\ \rho}(\eta^{\rho\sigma} + F^{\rho\sigma})L^{\nu}_{\ \sigma}\,. \tag{20.34}$$

In matrix notation,

$$\bar{\eta}' + \bar{F}' = L\,(\bar{\eta} + \bar{F})\,L^{\mathrm{T}}\,. \tag{20.35}$$

Since $\det L = \det L^{\mathrm{T}}$ and $(\det L)^2 = 1$, we can take determinants to conclude immediately that

$$\det\,(\,\bar{\eta}' + \bar{F}'\,) = \det\,(\,\bar{\eta} + \bar{F}\,)\,. \tag{20.36}$$

This establishes the Lorentz invariance of the Born–Infeld Lagrangian density.

One additional fact is easily derived. Since the determinant of a matrix does not change under transposition, we have

$$\det\bigl(\eta + F\bigr) = \det\bigl(\eta^{\mathrm{T}} + F^{\mathrm{T}}\bigr) = \det\bigl(\eta - F\bigr)\,. \tag{20.37}$$

We conclude that the Born–Infeld Lagrangian is an even function of F. The expression for the argument of the square root in (20.30) can be simplified, using $\eta = \eta^{-1}$ and $\det\eta = -1$:

$$-\det\bigl(\eta + F\bigr) = -\det\bigl(\eta(1 + \eta F)\bigr) = \det(1 + \eta F)\,. \tag{20.38}$$

Explicitly, the matrix $1 + \eta F$ is given by

$$1 + \eta F = \begin{pmatrix} 1 & E_x & E_y & E_z \\ E_x & 1 & B_z & -B_y \\ E_y & -B_z & 1 & B_x \\ E_z & B_y & -B_x & 1 \end{pmatrix}\,. \tag{20.39}$$

Each term of the determinant expansion is a product of terms, each of which contains precisely one element from each row and one element from each column. So the terms which are quadratic in the fields contain two elements from the diagonal. There are six ways of picking these two, and the corresponding terms give $-E^2 + B^2$, which is the expected answer. The quartic terms take a little more work to write out completely. In this case, one must not pick any term from the diagonal.

Our final topic in this section is the computation of the self-energy of a point charge in Born–Infeld theory. For this problem we can set $\vec{B} = 0$ in (20.28) and use the simplified Lagrangian density

$$\mathcal{L} = -b^2\sqrt{1 - \frac{E^2}{b^2}} + b^2\,. \tag{20.40}$$

As a first step, we calculate $\vec{D}$:

$$\vec{D} = \frac{\partial\mathcal{L}}{\partial\vec{E}} = \frac{\vec{E}}{\sqrt{1 - E^2/b^2}}\,. \tag{20.41}$$

We see that $\vec{E}$ and $\vec{D}$ point in the same direction. To solve for $\vec{E}$ in terms of $\vec{D}$, we first square the above equation and solve for E^2:

$$D^2 = \frac{E^2}{1 - E^2/b^2} \;\rightarrow\; E^2 = \frac{D^2}{1 + D^2/b^2}\,. \tag{20.42}$$

At this stage, by writing

$$E^2 = b^2\Bigl(\frac{D^2}{D^2 + b^2}\Bigr) = D^2\Bigl(\frac{b^2}{b^2 + D^2}\Bigr)\,, \tag{20.43}$$

we immediately see that

$$E \le b\,, \quad E \le D\,. \tag{20.44}$$

As expected, the electric field is bounded by b. Moreover, the magnitude E of the electric field is everywhere bounded by the magnitude D of $\vec{D}$. Note, however, that D can be arbitrarily large. In fact, it requires infinitely large D in order to get $E = b$.

To obtain $\vec{E}$, we take the "square root" of (20.42) by writing

$$\vec{E}\cdot\vec{E} = \frac{\vec{D}}{\sqrt{1+D^2/b^2}}\cdot\frac{\vec{D}}{\sqrt{1+D^2/b^2}}\,. \tag{20.45}$$

Since $\vec{E}$ points in the same direction as $\vec{D}$, the unique solution is

$$\vec{E} = \frac{\vec{D}}{\sqrt{1+D^2/b^2}}\,. \tag{20.46}$$

With $\vec{E}(\vec{D})$ at hand we are close to being able to calculate the energy density $\mathcal{H}(\vec{D})$. As a final ingredient, we need $\mathcal{L}$ written in terms of $\vec{D}$:

$$\mathcal{L} = -b^2\sqrt{1-\frac{D^2}{b^2(1+D^2/b^2)}} + b^2 = \frac{-b^2}{\sqrt{1+D^2/b^2}} + b^2\,. \tag{20.47}$$

The Hamiltonian is then

$$\mathcal{H} = \vec{E}\cdot\vec{D} - \mathcal{L} = \frac{D^2}{\sqrt{1+D^2/b^2}} + \frac{b^2}{\sqrt{1+D^2/b^2}} - b^2\,, \tag{20.48}$$

and the final result is

$$\mathcal{H} = b^2\sqrt{1+\frac{D^2}{b^2}} - b^2\,. \tag{20.49}$$

This is the Born–Infeld energy density when $\vec{B} = \vec{H} = 0$. You will confirm in Problem 20.4 that this result holds for Born–Infeld theory in an arbitrary number of dimensions.

Let us calculate the self-energy of a point charge in four-dimensional spacetime. In Maxwell theory the infinite self-energy comes about because the energy density is proportional to $ED = E^2$, and $E \sim r^{-2}$, where r is the distance to the charge. As a result, $d^3x\,E^2 \sim dr/r^2$, and the energy integral diverges for small r. In Born–Infeld theory we also find $D \sim r^{-2}$, but for large D the energy density (20.49) becomes

$$\mathcal{H} \simeq bD = E_{\rm crit}\,D, \quad \text{as } D\to\infty\,. \tag{20.50}$$

For large fields, the Maxwell energy density $u = \frac{1}{2}ED$ is replaced in Born–Infeld theory by $u = E_{\rm crit}\,D$. For large fields, the Born–Infeld energy grows linearly with D. As a result, $d^3x\,E_{\rm crit}D \sim d^3x\,r^{-2} \sim dr$, and the integral will converge. Let us now examine the details.

Suppose that we have a point charge Q. Because of spherical symmetry the field $\vec{D}$ is radial, and the equation $\nabla\cdot\vec{D} = \rho$ can be integrated over the volume bounded by a sphere of radius r to give

$$\oint_S \vec{D}\cdot d\vec{a} = Q \longrightarrow \vec{D}(\vec{r}) = \frac{Q}{4\pi r^2}\,\vec{e}_r\,, \tag{20.51}$$

where $\vec{e}_r$ is the unit vector in the radial direction. We must check that our solution satisfies $\nabla\times\vec{E}=0$. Because of (20.46), we know that $\vec{E}$ is of the form $\vec{E}=f(r)\vec{r}$. Such an electric field has zero curl:

Quick calculation 20.4 Show that $\nabla\times(f(r)\,\vec{r})=0$.

The energy U_Q of the point charge can now be calculated. Making use of (20.49) and (20.51), we write:

$$\begin{aligned} U_Q &= \int d^3x\,\mathcal{H} = b^2\int_0^\infty 4\pi r^2 dr\left(\sqrt{1+\left(\frac{Q}{4\pi b r^2}\right)^2}-1\right)\\ &= 4\pi\, b^2\int_0^\infty dr\left(\sqrt{r^4+\left(\frac{Q}{4\pi b}\right)^2}-r^2\right). \end{aligned} \tag{20.52}$$

Letting $r = x\sqrt{Q/4\pi b}$, we obtain

$$U_Q = \sqrt{\frac{b}{4\pi}}\;Q\sqrt{Q}\int_0^\infty dx(\sqrt{1+x^4}-x^2)\,. \tag{20.53}$$

This integral converges. The short distance problem has disappeared: the integrand is regular around $x=0$. We do not expect a problem at large x: the fields are small and the Born–Infeld energy is approximately equal to the Maxwell energy, for which no problem arises at large distances. Indeed,

$$\sqrt{1+x^4}-x^2 = \frac{1}{\sqrt{1+x^4}+x^2} < \frac{1}{2x^2}\,, \tag{20.54}$$

which is integrable around $x=\infty$. Doing the integral explicitly, one finds

$$\int_0^\infty dx(\sqrt{1+x^4}-x^2) = \frac{(\Gamma(1/4))^2}{6\sqrt{\pi}} \simeq \frac{(3.6256)^2}{6(1.7725)} \simeq 1.236\,. \tag{20.55}$$

Here $\Gamma(x)$ is the gamma function (see (3.50)). Back in (20.53), we find our final answer for the energy of a point charge in Born–Infeld theory:

$$U_Q = \frac{1}{4\pi}\frac{1}{3}\,(\Gamma(1/4))^2\,b^{1/2}\,Q^{3/2} \simeq \frac{1}{4\pi}\cdot 4.382\;b^{1/2}\,Q^{3/2}\,. \tag{20.56}$$

If b corresponds to the critical electric field in string theory, then $b=1/(2\pi\alpha')$. Moreover, $\sqrt{\alpha'}=\ell_s$, where ℓ_s is the string length. The self-energy of a point charge becomes

$$U_Q = \frac{1}{4\pi}\frac{1}{\sqrt{2\pi}}\frac{1}{3}\,(\Gamma(1/4))^2\frac{Q^2}{\ell_s\sqrt{Q}} \simeq \frac{1}{4\pi}\cdot 1.748\,\frac{Q^2}{\ell_s\sqrt{Q}}\,. \tag{20.57}$$

To appreciate this answer better, we compare it with the electrostatic energy of a charge distribution in Maxwell theory. If we assume that the charge is distributed uniformly over the volume of a ball of radius a, then the energy $U(Q)$ can be shown to be

$$U(Q) = \frac{1}{4\pi}\,\frac{3}{5}\,\frac{Q^2}{a}\,. \tag{20.58}$$

As $a \to 0$ the charge becomes point-like, and we obtain the expected infinite self-energy. We can think of a as a small smearing parameter that is introduced to make the self-energy finite. The classical electron radius, for example, is roughly the value of a for which $U(e)$ equals the rest energy of the electron. If we compare the Born–Infeld energy (20.57) with the Maxwell energy, we see that $\ell_s\sqrt{Q}$ plays the role of the smearing parameter. This quantity has the correct units because Q is dimensionless. It is interesting that the smearing parameter grows with the value of Q. The $Q^{3/2}$ dependence of the Born–Infeld energy U_Q is a sign of a nonlinear theory.

The above calculation of the self-energy of a static point charge does not seem to be directly applicable in string theory. On a space-filling D-brane open string endpoints are charged, but they also move. A tractable situation arises, however, when a semi-infinite open string ends on a Dp-brane, as you may examine in Problems 20.6 and 20.7. The string can then be static, and the Born–Infeld energy of the solution gives the energy associated with the semi-infinite string!

20.3 Born–Infeld theory and T-duality

In this chapter we have studied Born–Infeld electrodynamics in some detail. We were motivated by the existence of a critical electric field in string theory. In the present section we will show that our work on the T-duality properties of electric and magnetic fields on D-branes gives direct evidence that the dynamics of electromagnetic fields on D-branes is described by Born–Infeld theory.

For a static Dp-brane the product of its tension $T_p(g)$ and its volume gives the mass of the brane. Consider a world with one dimension curled up into a circle of radius R and $(p-1)$ dimensions curled up into some compact space of volume V_{p-1}. Now imagine a Dp-brane that wraps around the full set of p compact dimensions. The mass of this Dp-brane is

$$T_p(g)\,(2\pi R)\,V_{p-1}\,. \tag{20.59}$$

Under a T-duality transformation along the circle of radius R we obtain a D$(p-1)$-brane placed at some point on the dual circle, while the other $(p-1)$ directions along the brane world-volume still wrap around the compact space of volume V_{p-1}. The mass of this D-brane is given by

$$T_{p-1}(\tilde{g})\;V_{p-1}\,, \tag{20.60}$$

where $\tilde{g}$ is the string coupling in the dual picture (see (18.56)). Since T-duality is a physical equivalence, the two masses obtained above must be equal. Indeed, each D-brane is seen by a lower-dimensional observer as a point mass, and unless these masses are the same, the observer can tell that the physics has changed. Equating the values of the two masses, we find

$$T_{p-1}(\tilde{g}) = 2\pi R\,T_p(g)\,. \tag{20.61}$$

This is a relation between tensions of D-branes in *dual* pictures. $T_p(g)$ is the tension of a brane in the world with coupling constant g (the world with the circle of radius R), while $T_{p-1}(\tilde{g})$ is the tension of a brane in the world with coupling constant $\tilde{g}$ (the world with the circle of dual radius $\tilde{R}$). This relation between tensions holds generally, regardless of which coordinates the branes wrap around. The different looking relation between D-brane tensions in (18.57) compares tensions in the *same* picture, that is, with the same string coupling.

Let us now reconsider the configuration discussed in Section 19.3. We had two compact dimensions of radii R_2 and R_3 which formed a two-torus. One direction on the world-volume of a D$(p-1)$-brane was stretched along the diagonal of the torus. Assume, for simplicity, that the other $(p-2)$ directions wrap around a compact space of volume V_{p-2}. Let $\tilde{g}$ denote the string coupling in this picture. A T-duality transformation along the circle of radius R_3 gave us a Dp-brane with two directions wrapped around a torus with radii R_2, $\tilde{R}_3$ and a magnetic field. The other $(p-2)$ directions on the brane also wrap the space of volume V_{p-2}. Let g denote the string coupling in this picture. Since the circle that is dualized in the Dp-brane world is of radius $\tilde{R}_3$, equation (20.61) gives

$$T_{p-1}(\tilde{g}) = 2\pi \tilde{R}_3 \, T_p(g) \,. \tag{20.62}$$

For the static D$(p-1)$-brane that stretches along the torus diagonal, the Lagrangian is the negative of the brane rest energy. This energy is just the brane tension $T_{p-1}(\tilde{g})$ times the brane volume. The volume, in turn, is V_{p-2} times the length L_{diag} along the diagonal, so

$$L = -V_{p-2} L_{\text{diag}} \, T_{p-1}(\tilde{g}) = -V_{p-2} \sqrt{(2\pi R_2)^2 + (2\pi R_3)^2} \;\; T_{p-1}(\tilde{g}) \,. \tag{20.63}$$

We showed in (19.47) that $2\pi\alpha' B = -\tan\alpha$, where α is the angle that the diagonal in the torus makes with the horizontal direction. In fact, $\tan\alpha = R_3/R_2$, as indicated in (19.55). Therefore, the magnetic field is related to the ratio of the radii by

$$2\pi\alpha' B = -\frac{R_3}{R_2} \,. \tag{20.64}$$

Equations (20.62) and (20.64) allow us to rewrite (20.63) as follows:

$$L = -V_{p-2} \; (2\pi R_2) \sqrt{1 + (2\pi\alpha' B)^2} \;\; (2\pi \tilde{R}_3) \, T_p(g) \,. \tag{20.65}$$

Since $(2\pi R_2)(2\pi \tilde{R}_3)$ is the volume of the torus wrapped by the Dp-brane, we finally have

$$L = -V_p \; T_p(g) \sqrt{1 + (2\pi\alpha' B)^2} \;\; . \tag{20.66}$$

If $B = 0$, we recover the Lagrangian for a static Dp-brane.

To compare with the Born–Infeld Lagrangian, we consider (20.28), which for $\vec{E} = 0$ reduces to

$$\mathcal{L} = -b^2 \sqrt{1 + \frac{B^2}{b^2}} + b^2 \,. \tag{20.67}$$

Both the b^2 multiplying the square root and the b^2 added to $\mathcal{L}$ were introduced for reasons that are no longer relevant. The additive contribution, for example, was originally included to cancel the constant term in $\mathcal{L}$. Now the constant term is needed to represent the rest energy of the D-brane in the absence of electromagnetic fields. The b^2 multiplying the square root was originally included to give a standard normalization to the Maxwell action. Now the overall normalization of the action is fixed in (20.66). We cannot even rescale the gauge potentials. The normalization of the gauge field was fixed when we wrote the coupling (16.54) to the open string endpoints.

As a result, the value $b = 1/(2\pi\alpha')$, that we read by comparing the terms inside the square roots in (20.66) and (20.67), is only to be used inside the square root. Therefore, on account of (20.29), the D-brane Lagrangian (20.66) is consistent with the Born–Infeld Lagrangian density

$$\mathcal{L} = -\,T_p(g)\,\sqrt{-\det(\eta_{mn} + 2\pi\alpha' F_{mn})}\ . \tag{20.68}$$

Here m, n are indices on the world-volume of the Dp-brane. The volume factor V_p was removed in order to write the density $\mathcal{L}$. This Lagrangian density describes the behavior of electromagnetic fields on D-branes.

Quick calculation 20.5 Assume that $F_{23} = B$ is the only nonvanishing magnetic field component. Evaluate the Lagrangian density (20.68) for an arbitrary Dp-brane ($p \geq 3$), and verify that your result is consistent with (20.66).

The Lagrangian density (20.68) allows us to calculate the electromagnetic fields that arise from open string endpoints, which, in our conventions, carry charges of unit magnitude. For a static oriented open string that ends at $\vec{x}_0$, for example, the charge density is given by $\rho(\vec{x}) = \delta(\vec{x} - \vec{x}_0)$, where $\vec{x}$ represents the spatial coordinate on the D-brane. The equation for $\vec{D}$ then reads

$$\nabla \cdot \vec{D} = \delta(\vec{x} - \vec{x}_0)\,, \qquad \vec{D} = \frac{\partial \mathcal{L}}{\partial \vec{E}}\,. \tag{20.69}$$

A particular version of this equation is solved in Problem 20.7. The solution describes a string ending on a D-brane.

The T-duality analysis of electric fields in Section 19.2 supports the identification of (20.68) as the Lagrangian density. There we had a circle of radius $\tilde{R}$ and a D$(p-1)$-brane moving with velocity v along this circle. How do we write a Lagrangian for such a D-brane? For a point particle of mass m, the Lagrangian is $(-m)$ times the relativistic factor $\sqrt{1 - v^2/c^2}$, as shown in (5.8). For a string, the Lagrangian density is (minus) the rest energy $(-T_0 ds)$ of a piece of string times the analogous relativistic factor (see (6.89)). For a moving D$(p-1)$-brane we write

$$L = -V_{p-1}\,T_{p-1}(\tilde{g})\,\sqrt{1 - \frac{v^2}{c^2}}\,. \tag{20.70}$$

We now re-express this Lagrangian in terms of the variables of the T-dual picture, where we have a Dp-brane with an electric field along a circle of radius R and the string coupling is g. The value of the electric field is related to the brane velocity by $v/c = \beta = 2\pi\alpha' E$ (see (19.34)). We thus find

$$L = -V_{p-1}\,(2\pi R)\,T_p(g)\,\sqrt{1-(2\pi\alpha' E)^2}\ , \tag{20.71}$$

where we also made use of (20.61). Finally, the product of $2\pi R$ and V_{p-1} gives the total volume V_p of the Dp-brane:

$$L = -V_p\ T_p(g)\ \sqrt{1-(2\pi\alpha' E)^2}\ . \tag{20.72}$$

This Lagrangian is correctly reproduced by (20.68) when we have a constant electric field and zero magnetic field. This can be easily checked by letting $F_{01} = -E_x$ be the only nonvanishing electric field component. We have thus obtained further evidence for the correctness of (20.68).

Problems

Problem 20.1 Energy functional in nonlinear electrodynamics.

(a) According to equation (20.21), the total energy U in a fixed volume V is

$$U = \int_V \Big(\vec{D}\cdot\vec{E} - \mathcal{L}\Big) d^3x\,. \tag{1}$$

Prove that, in four-dimensional spacetime,

$$\frac{dU}{dt} = \int_V \Big(\vec{E}\cdot\frac{\partial\vec{D}}{\partial t} + \vec{H}\cdot\frac{\partial\vec{B}}{\partial t}\Big) d^3x\,. \tag{2}$$

(b) Use equations (20.1) and (20.2) to show that, in the absence of sources,

$$\frac{dU}{dt} = -\int_S (\vec{E}\times\vec{H})\cdot d\vec{a}\,, \tag{3}$$

where S is the surface that bounds the volume V. This shows that U is constant if the electromagnetic fields vanish at the boundary and thus carry no energy out of V. U is a conserved energy. Indeed, you may recognize $\vec{E}\times\vec{H}$ as the Poynting vector, which represents local energy flow per unit area per unit time.

Problem 20.2 Capacitance in Born–Infeld theory.

The capacitance C of a two-conductor configuration is defined as $Q = CV$, where the conductors have charges Q and $-Q$, respectively, and V is the potential difference between them.

(a) Let $C_{\rm M}$ denote the capacitance in Maxwell theory. Explain why $C_{\rm M}$ is a constant, independent of Q and V.

(b) Consider a two-conductor configuration, with charges Q and $-Q$, and let $\vec{E}_{\rm M}(\vec{x})$ be the electric field between the conductors in Maxwell theory. Prove that whenever $\nabla E_{\rm M}^2$ points in the direction of $\vec{E}_{\rm M}$, the field $\vec{D}(\vec{x})$ in Born–Infeld theory, for the same charged configuration, is given by $\vec{D}(\vec{x}) = \vec{E}_{\rm M}(\vec{x})$. Prove that in such cases the Born–Infeld capacitance is always greater than the Maxwell capacitance.

(c) Consider a parallel plate capacitor of area A and plate separation d. Show that the Born–Infeld capacitance $C(V)$ is

$$C(V) = \frac{C_{\rm M}}{\sqrt{1-(V/V_c)^2}}\,,$$

where $C_{\rm M} = A/d$ is the Maxwell capacitance and $V_c = bd$. What is the interpretation of V_c?

Problem 20.3 Dual field strength and the Lorentz invariant p.

(a) Show that the replacement $\tilde{F}^{\mu\nu} \to F^{\mu\nu}$ corresponds to $\vec{E} \to -\vec{B}$ and $\vec{B} \to \vec{E}$. Verify that these are the duality transformations of electromagnetism (Section 17.1).

(b) Show that the Lorentz invariant $p = -\frac{1}{4}\tilde{F}^{\mu\nu}F_{\mu\nu}$ can be written as a total derivative: $p = -\frac{1}{4}\partial_\mu(\epsilon^{\mu\nu\rho\sigma}F_{\nu\rho}A_\sigma)$.

Problem 20.4 Electric fields and point charges in higher dimensions.

(a) Examine the Lagrangian density (20.29) when there is only an electric field, $F_{0i} = -E_i$, in a world with $D = d+1$ spacetime dimensions. Calculate explicitly the determinant, and show that it takes the form given in (20.40), with $E^2 = \vec{E}\cdot\vec{E}$.

(b) The calculation of the determinant in (a) can be simplified. We have proven the Lorentz invariance of the Born–Infeld Lagrangian density and, therefore, its rotational invariance. Imagine calculating the Lagrangian density at the origin, using a set of axes for which $\vec{E}$ is aligned along the first spatial coordinate. Show how the almost trivial computation in this frame can be used to anticipate the answer obtained in (a).

(c) The energy U_Q of a point charge Q in a D-dimensional spacetime is proportional to Q^δ, where δ is a constant. Calculate δ.

Problem 20.5 Calculating the Born–Infeld Hamiltonian.

Consider the full Born–Infeld Lagrangian density $\mathcal{L}$, with $b = 1$:

$$\mathcal{L} = -\sqrt{1 - E^2 + B^2 - (\vec{E}\cdot\vec{B})^2} + 1\,. \tag{1}$$

(a) Show that

$$\vec{D} = \frac{\vec{E} + (\vec{E}\cdot\vec{B})\vec{B}}{\sqrt{1 + B^2 - E^2 - (\vec{E}\cdot\vec{B})^2}}\,. \tag{2}$$

(b) Now the challenge is to solve for $\vec{E}$ in terms of $\vec{D}$. This requires quite a bit of trickery in vector algebra. As a first step note that equation (2) is of the form $f(\vec{E}, \vec{B})\,\vec{D} = \vec{E} + (\vec{E}\cdot\vec{B})\vec{B}$, where f is a scalar function. Show that

$$\vec{E} = \frac{f(\vec{E}, \vec{B})}{1+B^2}\left(\vec{D} - (\vec{D}\times\vec{B})\times\vec{B}\right). \tag{3}$$

The function f contains the combination $E^2 + (\vec{E}\cdot\vec{B})^2$. To write this combination in terms of $\vec{D}$ and $\vec{B}$, examine the values of D^2 and $(\vec{D}\times\vec{B})^2$. Now prove that

$$\vec{E} = \frac{\vec{D} - (\vec{D}\times\vec{B})\times\vec{B}}{\sqrt{1+B^2+D^2+(\vec{D}\times\vec{B})^2}}\,. \tag{4}$$

(c) As you are now equipped with the value of $\vec{E}$, the calculation of the Hamiltonian is relatively simple. Use (20.21) to show that

$$\mathcal{H} = \sqrt{1+B^2+D^2+(\vec{D}\times\vec{B})^2} \;-\; 1\,. \tag{4}$$

Problem 20.6 String ending on a D-brane in Born–Infeld theory: Part 1.

The dynamics of electromagnetic fields on a Dp-brane is governed by (20.68). But we know that there are also $(d-p)$ massless scalar fields living on the world-volume of a Dp-brane, whose excitations represent transverse displacements of the brane (recall the discussion below (15.37)). Consider the scalar field X associated with displacements along an x axis normal to the brane. The Born–Infeld action can be generalized to describe both the electromagnetic and X fields:

$$\mathcal{L} = -T_p\sqrt{-\det\left(\eta_{mn} + 2\pi\alpha' F_{mn} + \partial_m X\,\partial_n X\right)}\;. \tag{1}$$

The field X here is a function of the coordinates x^m on the brane (X is *not* a string coordinate). The value $X(x^m)$ is the x coordinate of the point on the brane with coordinates x^m.

(a) Suppose there are no electromagnetic fields on the Dp-brane, but we let $X = vt$, attempting to represent the motion of the D-brane along the normal direction x with velocity v. Show that $\mathcal{L} = -T_p\sqrt{1-v^2/c^2}$, which is the expected Lagrangian density for a moving D-brane. This result supports the interpretation of X given above.

We want to evaluate the Lagrangian density (1) in the case where there is only an electric field $\vec{E}$. For notational convenience introduce $\vec{\mathcal{E}} = 2\pi\alpha'\vec{E}$.

(b) To simplify the evaluation of the determinant we use symmetry arguments, as in part (b) of Problem 20.4. Choose a set of axes so that the electric field points along the first direction: only $\mathcal{E}_1$ is nonzero. There is another vector in this problem: the gradient ∇X of the field X. We choose the axes in such a way that this vector lies in the plane formed by the first and second directions, so that only $(\nabla X)_1 \equiv X_1$ and $(\nabla X)_2 \equiv X_2$ are nonzero. Show that under these conditions the determinant in (1) gives

$$\det(\eta_{mn} + \cdots) = -1 + \dot{X}^2 - X_1^2 - X_2^2 + \mathcal{E}_1^2 + \mathcal{E}_1^2\,X_2^2\,. \tag{2}$$

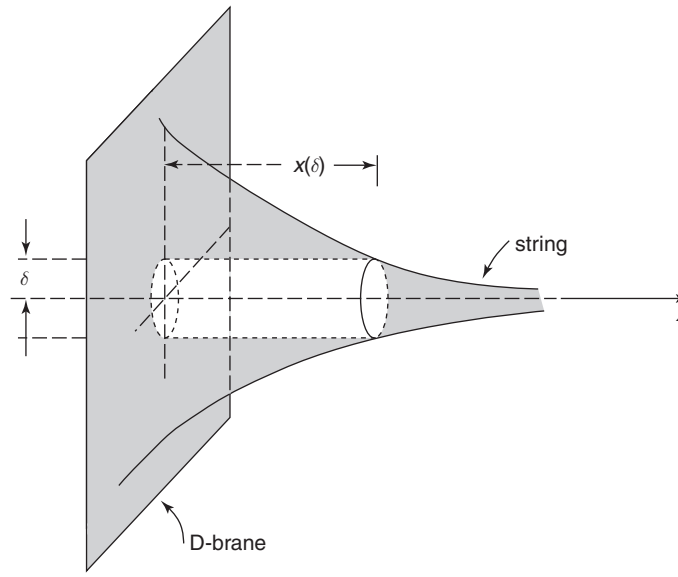

Fig. 20.1 **Problem 20.7: a string that ends on a D-brane is represented in Born–Infeld theory as a solution in which the brane itself is deformed.**

(c) The rotational invariance of the determinant implies that it can be written in terms of dot products of the vectors $\vec{\mathcal{E}}$ and ∇X and in terms of the rotation scalar $\dot{X}$. Use this requirement to deduce that

$$\mathcal{L} = -\,T_p\sqrt{(1-\mathcal{E}^2)(1+(\nabla X)^2)+(\vec{\mathcal{E}}\cdot\nabla X)^2-\dot{X}^2}\ . \tag{3}$$

Problem 20.7 String ending on a D-brane in Born–Infeld theory: Part 2.

The description of a string ending on a Dp-brane in Born–Infeld theory involves an electric field on the brane, due to the charged endpoint, and an excitation of the scalar field X that represents brane displacements in the direction along the stretched string (see Figure 20.1). This direction is assumed to be normal to the brane. We can therefore use the Lagrangian density (3), obtained in Problem 20.6:

$$\mathcal{L} = -\,T_p\sqrt{(1-\mathcal{E}^2)(1+(\nabla X)^2)+(\vec{\mathcal{E}}\cdot\nabla X)^2-\dot{X}^2}\ . \tag{1}$$

Here the electric field is: $\vec{\mathcal{E}} = 2\pi\alpha'\vec{E} = \nabla\mathcal{A}_0 - \partial_t\vec{\mathcal{A}}$, where $\mathcal{A}_\mu \equiv 2\pi\alpha' A_\mu$.

(a) Explain why the equations of motion that follow from the variation of $\vec{\mathcal{A}}$ are satisfied if all fields are time independent.

We will therefore assume that all fields are indeed time independent, and, moreover, we will set $\vec{\mathcal{A}} = 0$, which is consistent with our assumption that there are no magnetic fields F_{ij}. In this case, the Lagrangian density becomes

$$\mathcal{L} = -\,T_p\sqrt{(1-(\nabla\mathcal{A}_0)^2)(1+(\nabla X)^2)+(\nabla\mathcal{A}_0\cdot\nabla X)^2}\ . \tag{2}$$

(b) Derive the equations of motion that arise from the variations δX and $\delta\mathcal{A}_0$ of the action associated with the Lagrangian density (2). Write them in the form $\nabla\cdot[\ldots]=0$. Show that both equations are satisfied if

$$\vec{\mathcal{E}} = \nabla\mathcal{A}_0 = \pm\nabla X\,, \quad \nabla^2\mathcal{A}_0 = \nabla^2 X = 0\,, \tag{3}$$

for either choice of sign.

(c) Show that when $\vec{\mathcal{E}} = \nabla\mathcal{A}_0 = \pm\nabla X$ (equation (3)) holds we have

$$\vec{D} = 2\pi\alpha'\,T_p\,\vec{\mathcal{E}}\,, \tag{4}$$

and the energy $U = \int d^p x(\vec{D}\cdot\vec{E}-\mathcal{L})$ is given by

$$U = T_p\int d^p x\big(1+\vec{\mathcal{E}}\cdot\vec{\mathcal{E}}\big)\,. \tag{5}$$

The first term in U represents the rest energy of the Dp-brane. The second term gives the energy associated with the string ending on the D-brane, as we discuss next.

Let r denote radial distance on the D-brane. The solution for a string *ending* at $r=0$ is obtained solving $\nabla\cdot\vec{D} = +\delta(r)$ (see (20.69)). The solution for $\vec{D}$ (and for $\vec{\mathcal{E}}$) is spherically symmetric; so is the solution for X, which is of the form $X(r)$. Additionally, we assume that $p\geq 3$, in which case we can require that $X(\infty)=0$ (explain why!). The sign in $\vec{\mathcal{E}} = \pm\nabla X$ determines whether the string stretches along the positive x axis (as in the figure) or along the negative x axis. Which sign do you need to get the option in the figure?

(d) Show that the energy $U_s = T_p\int d^p x\,\vec{\mathcal{E}}\cdot\vec{\mathcal{E}}$ is infinite.

(e) To interpret the infinite value of U_s, consider the region on the D-brane with $r>\delta$, and let $U_s(\delta)$ denote the energy contained in this region. Show that

$$U_s(\delta) = T_p\int_{r>\delta} d^p x\ \nabla X\cdot\nabla X = T_p|X(\delta)|\ \cdot\text{Flux of }\vec{\mathcal{E}}\text{ across }S^{p-1}(\delta)\,, \tag{6}$$

where $S^{p-1}(\delta)$ denotes the $(p-1)$-dimensional sphere of radius δ. Conclude finally that

$$U_s(\delta) = \frac{1}{2\pi\alpha'}\,|X(\delta)|\,. \tag{7}$$

Since $1/(2\pi\alpha')$ is the string tension, this confirms that $U_s(\delta)$ is the energy of the piece of string that stretches from $x=0$ up to $x=|X(\delta)|$. As $\delta\to 0$ the energy diverges because this is the energy of an infinitely long string.

21 String theory and particle physics

Configurations of intersecting D6-branes in type IIA superstring theory define string models of particle physics. Open string states supported at the intersection of the branes naturally give chiral fermions, a key ingredient of the Standard Model. If orientifold planes are included, the models can display the massless spectrum of gauge bosons and chiral fermions of the Standard Model. Compactification moduli are adjustable parameters that give rise to undesirable massless scalars and must be stabilized. Flux compactifications achieve moduli stabilization and give rise to an extremely large landscape of string vacua. The existence of vacua in which the vacuum energy matches the presently observed value becomes statistically plausible.

21.1 Intersecting D6-branes

In this section we consider a D-brane configuration that has a set of features that make it a good starting point for the construction of a string model of particle physics. Since we need fermions, we use a ten-dimensional superstring theory. In this theory, six of the ten dimensions, $x^4, \ldots, x^9$, are taken to form a small compact space of finite volume. This is necessary in order to have an effectively four-dimensional spacetime (with coordinates x^0, x^1, x^2, and x^3). The compact space is as simple as possible: each dimension is turned into a circle, so that the resulting space is a six-dimensional torus T^6. We will assume that all circles have the same radius R, so we are taking $x^i \sim x^i + 2\pi R$ for $i = 4, \ldots, 9$.

In order to obtain an effective four-dimensional Yang–Mills theory, we need D-branes that have at least three spatial directions to stretch along the spatial coordinates x^1, x^2, and x^3 of the effective spacetime. Therefore, we will use Dp-branes with $p \geq 3$. In fact, in this section we will work with D6-branes in type IIA superstring theory. This is, of course, just one choice among many that we could have made in order to construct a model. In addition, we will let the D6-branes intersect. When two D-branes intersect, one discovers a sector of open strings that stretch from one brane to the other and are localized near the intersection. Under certain circumstances, such strings give rise to matter fields with the properties of Standard Model fermions. We will examine this point in detail in Sections 21.3 and 21.4. Our main goal in the present section is to understand the geometry of intersecting D-branes.

Let us describe two D6-branes that intersect on the torus T^6 introduced above. For a D6-brane, of the nine spatial directions in the ten-dimensional spacetime, three directions are Dirichlet and six are Neumann. Three of the Neumann directions must be x^1, x^2, and x^3, as mentioned above, and the other three lie on T^6. The three Dirichlet directions are also on T^6. The position of the D6-brane can be specified by giving the values of the coordinates along the three Dirichlet directions. Thus, let the first D6-brane be defined by

$$\text{D6-brane \#1:} \quad x^5 = x^7 = x^9 = 0, \tag{21.1}$$

and the second D6-brane be defined by

$$\text{D6-brane \#2:} \quad x^4 = x^6 = x^8 = 0. \tag{21.2}$$

A point belongs to the intersection if it belongs to both D6-branes. The conditions (21.1) must hold for a point to be on the first D6-brane, and the conditions (21.2) must hold for a point to be on the second D6-brane. As a result, a point belongs to the intersection if

$$\text{intersection conditions}: x^4 = x^5 = x^6 = x^7 = x^8 = x^9 = 0. \tag{21.3}$$

Since the torus T^6 is spanned by the above coordinates, the D6-branes intersect on a *point* on the torus, the point $(0, 0, 0,\ 0, 0, 0)$. On the spacetime, however, the intersection of the D6-branes is the set of points

$$\text{intersection set}: (x^0, x^1, x^2, x^3,\ 0, 0, 0,\ 0, 0, 0), \quad \text{with} \quad x^0, x^1, x^2, x^3 \in \mathbb{R}. \tag{21.4}$$

The intersection of the D6-branes fills the effective four-dimensional spacetime.

A simple way to visualize the intersection of the D6-branes is to describe the six-torus T^6 in terms of three square two-tori T^2. One writes $T^6 = T^2 \times T^2 \times T^2$, in the same sense that one writes $\mathbb{R}^3 = \mathbb{R} \times \mathbb{R} \times \mathbb{R}$. Indeed, a six-torus is equivalent to a two-torus in the x^4 and x^5 directions, a two-torus in the x^6 and x^7 directions, and a two-torus in the x^8 and x^9 directions. These three two-tori are shown in Figure 21.1. On each T^2, each brane appears as a line; in fact, brane #1 appears as a horizontal line on account of (21.1), and brane #2 appears as a vertical line on account of (21.2). On each T^2, these straight lines are in fact circles because their endpoints are identified. The two D6-branes intersect at a point on T^6 because these lines intersect at a point on *each* T^2. Note that the full intersection is characterized by three intersection angles, one for each T^2. In the present case, these angles are all equal to $\pi/2$. The figure also shows a string in the sector [12]. On each of the T^2 we see a projection of the string as it stretches from the first brane to the second brane.

In the same way as we did for other D-brane configurations, we can use a table to describe the two D-branes and the boundary conditions for the various open string sectors. The result is given in Table 21.1. It is noteworthy that in the [12] and [21] sectors, the string coordinates along the torus are of DN or ND types. This happens because the branes intersect orthogonally on each T^2. We have chosen axes such that any coordinate on T^6 lies along one brane and is orthogonal to the other one. For arbitrary intersection angles, the boundary conditions for the string coordinates are slightly more complicated (Problem 15.7).

We are interested in D6-brane configurations obtained by allowing more general intersection angles between the lines that represent the D-branes on each T^2. For this purpose,

Table 21.1 A configuration of two intersecting D6-branes

Coordinate	$x^{1,2,3}$	x^4	x^5	x^6	x^7	x^8	x^9
D6 #1	–	–	•	–	•	–	•
D6 #2	–	•	–	•	–	•	–
Sector [1,1]	NN	NN	DD	NN	DD	NN	DD
Sector [2,2]	NN	DD	NN	DD	NN	DD	NN
Sector [1,2]	NN	ND	DN	ND	DN	ND	DN
Sector [2,1]	NN	DN	ND	DN	ND	DN	ND

Note: We denote by – coordinates along which the D-branes are stretched and by • coordinates normal to the branes. The last four rows give the boundary conditions for string coordinates in the four possible open string sectors.

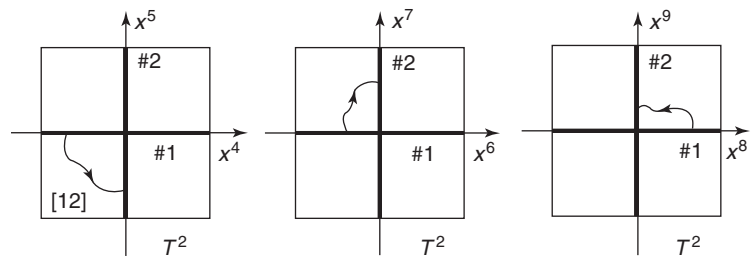

Fig. 21.1 **Two D6-branes on a x^6, which is presented as the product of three two-tori x^2. Brane #1 stretches along x^4, x^6, and x^8 and appears as a thick horizontal line. Brane #2 stretches along x^5, x^7, and x^9 and appears as a thick vertical line. A string stretching from the first to the second brane is also shown.**

we must first understand how two lines intersect on a two-torus. The lines that we will examine are in fact closed straight lines: the projections of the branes into the torus form circles, or line segments with identified endpoints. This is physically reasonable, the tension of a brane forces them to be straight. Moreover, unless they are closed, they will have infinite length, which requires infinite rest energy.

We can work with a torus defined on the (x, y) plane by the identifications $x \sim x + 1$ and $y \sim y + 1$, where by convention we have chosen the length between identified points to equal unity. The torus can be viewed as the unit square $0 \leq x, y \leq 1$ with boundary identifications. We can use two relatively prime integers (m, n) to describe an *oriented* closed line on this torus: the line is constructed on the (x, y) plane as an oriented straight segment from the origin $(0, 0)$ up to the point (m, n). The identifications can then be used to exhibit this segment fully on the square $0 \leq x, y, \leq 1$. Note that $(m, n) \sim (0, 0)$ under the identifications, so the segment is a closed line on the torus. Since m and n are relatively prime, the segment on the plane does not encounter a point with integer coordinates between its endpoints. This means that the line on the torus does not close before the segment ends. If m and n are not relatively prime, then their greatest common divisor represents the number

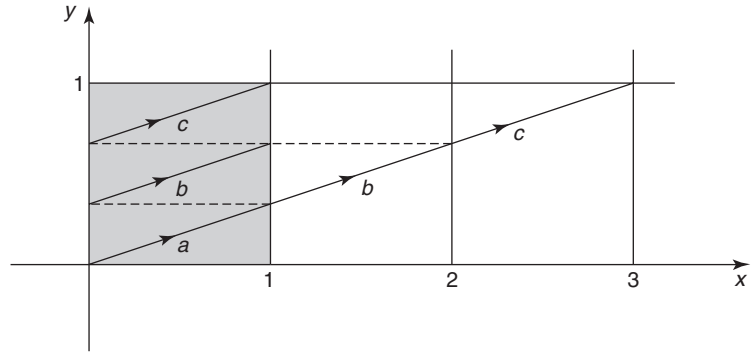

Fig. 21.2 **The segment joining (0, 0) to (3, 1) and the corresponding closed line $\ell = (3, 1)$ on the (shaded) torus. The segment goes through three copies of the fundamental domain, so it appears on the fundamental domain as three segments, denoted by *a*, *b*, and *c*.**

of times that the line is wrapped on the torus. We will focus on lines for which m and n are relatively prime. If one of the two integers is zero, then the other must be either plus or minus one, otherwise the curve will be multiply wrapped. An example showing the line (3, 1) is given in Figure 21.2.

Now consider a line $\ell_1 = (m_1, n_1)$ and a different line $\ell_2 = (m_2, n_2)$. The question is: how many times do these lines intersect on the torus? The answer is simple to state (and we will discuss its origins a little later on): the intersection number $\#(\ell_1, \ell_2)$ is

$$\#(\ell_1, \ell_2) = m_1 n_2 - m_2 n_1 = \det \begin{pmatrix} m_1 & n_1 \\ m_2 & n_2 \end{pmatrix}. \tag{21.5}$$

This intersection number can be interpreted as the z component of the vector cross product of $(m_1, n_1, 0)$ and $(m_2, n_2, 0)$. This means that the magnitude of the intersection number coincides with the area of the parallelogram defined by the vectors ℓ_1 and ℓ_2. As defined, the intersection number is antisymmetric under the exchange of the two lines: $\#(\ell_1, \ell_2) = -\#(\ell_2, \ell_1)$. The sign associated with the intersection has a meaning because we are dealing with oriented lines. When $\#(\ell_1, \ell_2) > 0$, the oriented line ℓ_1 aligns with the oriented line ℓ_2 after a *counterclockwise* rotation with angle less than π. When $\#(\ell_1, \ell_2) < 0$, the alignment occurs after a *clockwise* rotation with angle less than π. For example, the intersection number of $\ell_1 = (1, 0)$ with $\ell_2 = (0, 1)$ is equal to

$$\#(\ell_1, \ell_2) = 1 \times 1 - 0 \times 0 = +1. \tag{21.6}$$

Consistent with the plus sign, a counterclockwise rotation of $\ell_1 = (1, 0)$, with angle $\pi/2$, gives $\ell_2 = (0, 1)$. Since we are dealing with straight lines, the intersection angle at each intersection point is the same and the rotation needed to align the oriented lines is of the same type.

Quick calculation 21.1 Consider $\ell_1 = (-1, 1)$ and $\ell_2 = (1, 1)$. Calculate $\#(\ell_1, \ell_2)$, identify the intersection points on the torus, and confirm that the sign of the intersection number is in accord with expectations.

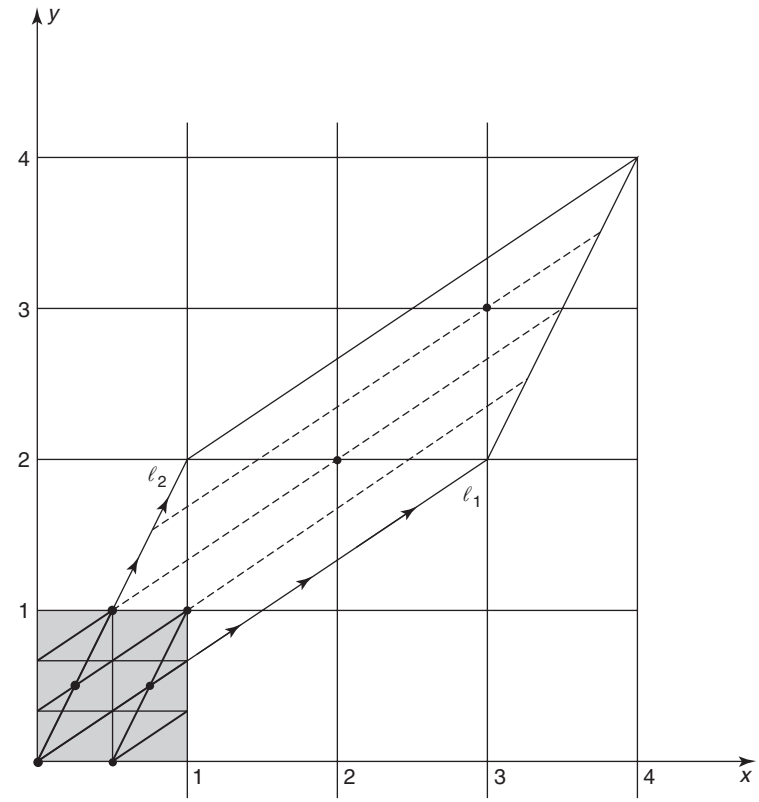

Fig. 21.3 **The lines $\ell_1 = (3, 2)$ and $\ell_2 = (1, 2)$ intersect four times on the (shaded) two-torus. The cell $\mathcal{C}$ spanned by ℓ_1 and ℓ_2 is a parallelogram with area four.**

In Figure 21.3 we show the intersection of the lines $\ell_1 = (3, 2)$ and $\ell_2 = (1, 2)$. When drawn on the square $0 \leq x, y \leq 1$, the line ℓ_1 appears as four segments, and the line ℓ_2 appears as two segments. We find two intersection points in the interior of the square. There is one intersection point on the horizontal sides (it is identified on the two sides). And there is one intersection point at the corners, all of which are identified. This gives a total of four intersection points. The intersection number can be calculated directly using (21.5), and we get the expected answer: $\#(\ell_1, \ell_2) = 3 \times 2 - 1 \times 2 = 4$.

The formula (21.5) for the intersection number gives zero for the intersection of a line with itself. A line coincides fully with itself, so what is the meaning of the zero? The meaning is topological: given two coincident lines on a torus, we can displace one of them a little and obtain a situation in which the lines no longer intersect. Consider the line $(1, 0)$, for example. A slightly displaced version of this closed line is $0 \leq x \leq 1, \ y = \epsilon$, with ϵ a small positive number. This new closed line has no point in common with the closed line $(1, 0)$. The intersection number is a topological quantity; roughly, the intersection number of two lines does not change under small deformations of the lines. That is why the intersection number of coincident lines is zero.

Quick calculation 21.2 Display on a torus the line $\ell = (1, 1)$ and a second line, parallel to ℓ, which is obtained by a small displacement of ℓ.

We now briefly explain the origin of (21.5). Consider the (x, y) plane with the earlier unit identifications $x \sim x + 1$ and $y \sim y + 1$, and think of the lattice of points with integer coordinates. The unit cell in this lattice is spanned by the vectors $(1, 0)$ and $(0, 1)$. The identifications on the unit cell give the unit torus. The cell $\mathcal{C}$ spanned by the vectors ℓ_1 and ℓ_2 is a parallelogram with some integer area I. We can use $\mathcal{C}$ to construct a big torus $\hat{\mathcal{C}}$ by gluing together the parallel sides of $\mathcal{C}$. If we denote by $\vec{x} = (x, y)$ an arbitrary point on the plane, the big torus $\hat{\mathcal{C}}$ arises from the identifications $\vec{x} \sim \vec{x} + \ell_1$ and $\vec{x} \sim \vec{x} + \ell_2$. We want to show that the area I of $\hat{\mathcal{C}}$ is the number of times that the closed lines ℓ_1 and ℓ_2 intersect on the unit torus. It is a known fact about lattices that the torus $\hat{\mathcal{C}}$ contains I copies of each point on the unit torus. Let us now count the total number of copies that arise from the set of intersection points on the unit torus. For this we must find all the intersections between all the copies of ℓ_1 and all the copies of ℓ_2. The torus $\hat{\mathcal{C}}$ contains I copies of the line ℓ_1: one copy is provided by the edge ℓ_1 of $\mathcal{C}$ (and the other parallel edge), and the other $I - 1$ copies are segments parallel to ℓ_1 that go through the interior of $\mathcal{C}$. Similarly, $\hat{\mathcal{C}}$ contains I copies of the line ℓ_2. On $\hat{\mathcal{C}}$, each copy of ℓ_1 intersects once with each copy of ℓ_2, so we have a total of $I \times I$ distinct intersection points on $\hat{\mathcal{C}}$. It follows that there are $I^2/I = I$ distinct intersection points on the unit torus. This is what we wanted to prove.

Quick calculation 21.3 Explain why the interior of $\mathcal{C}$ contains $I - 1$ points with integer coordinates. Explain also why the copies of ℓ_1 and ℓ_2 on the interior of $\mathcal{C}$ go through the points with integer coordinates.

Quick calculation 21.4 Find the coordinates of the intersection points between the lines $\ell_1 = (3, 2)$ and $\ell_2 = (1, 2)$ (Figure 21.3). Express your answers using points on the unit cell $0 \leq x, y \leq 1$.

We can now return to our D6-branes, which have three directions wrapped on T^6. Each wrapping is specified by three lines (ℓ_1, ℓ_2, ℓ_3). The line ℓ_i represents the direction of the D-brane on the ith T^2. The first D6-brane in our earlier example is now described by $\ell_1 = \ell_2 = \ell_3 = (1, 0)$. The second brane is described by $\ell_1 = \ell_2 = \ell_3 = (0, 1)$. When we specify a D6-brane using three oriented lines, we are actually giving an orientation to the three-dimensional subspace of the D6-brane that lies on the T^6. The orientation of a space of dimension k is defined by a choice of an ordered set of k linearly independent tangent vectors. The two-dimensional (x, y) plane, for example, can be oriented using the ordered pair of tangent vectors $((1, 0), (0, 1))$. Intuitively, this defines a circulation on the plane; the circulation is the direction in which one must rotate the first vector (with an angle smaller than π) to align it with the second vector. In more elementary descriptions, the orientation of a surface inside $\mathbb{R}^3$ is defined by an oriented direction normal to the surface. The two descriptions are related: the cross product of the ordered tangent vectors on the surface points in the direction of the oriented normal. For the three-dimensional subspace of a D6-brane lying on T^6, the set of ordered vectors is (ℓ_1, ℓ_2, ℓ_3).

Consider now a general situation where we have two D6-branes, a and b, each of which has three directions wrapped around T^6. Such a configuration is specified by

$$\begin{aligned} &\text{D6-brane } a\text{: } (\ell_1^{(a)}, \ell_2^{(a)}, \ell_3^{(a)}), \\ &\text{D6-brane } b\text{: } (\ell_1^{(b)}, \ell_2^{(b)}, \ell_3^{(b)}). \end{aligned} \tag{21.7}$$

How many times do these D6-branes intersect on T^6? They intersect a number of times I_{ab} equal to the product of the intersection numbers of the corresponding lines on each of the three two-tori. Indeed, the general intersection point on T^6 is obtained by choosing one intersection point from each of the three two-tori. We thus have

$$I_{ab} = \#(\ell_1^{(a)}, \ell_1^{(b)}) \cdot \#(\ell_2^{(a)}, \ell_2^{(b)}) \cdot \#(\ell_3^{(a)}, \ell_3^{(b)}) = \prod_{i=1}^{3} \#(\ell_i^{(a)}, \ell_i^{(b)}). \tag{21.8}$$

If we let $\ell_i^{(a)} = (m_i^a, n_i^a)$ and $\ell_i^{(b)} = (m_i^b, n_i^b)$, then

$$I_{ab} = \prod_{i=1}^{3} (m_i^a\, n_i^b - m_i^b n_i^a). \tag{21.9}$$

The number I_{ab} carries a sign, the interpretation of which is straightforward. Since we have three intersections, there are three rotations that are needed to align brane a with brane b. The sign of I_{ab} is negative if the number of clockwise rotations is odd, and it is positive if the number of clockwise rotations is even. I_{ab} is the intersection number of the two oriented D6-branes. In particle physics models constructed with intersecting D6-branes, four-dimensional chiral fermions arise from open strings localized near each intersection. Even more, the sign of the intersection number determines the orientation of the open string which represents the chiral fermions. We will consider these matters in some detail in Sections 21.3 and 21.4.

If $I_{ab} = 0$, then at least one of the three intersection numbers $\#(\ell_i^{(a)}, \ell_i^{(b)})$, $i = 1, 2, 3$, must be equal to zero. Suppose that the intersection number vanishes only for $i = 1$. Then $\ell_1^{(a)}$ and $\ell_1^{(b)}$ must be either parallel or antiparallel. Since the lines we are considering are defined by relatively prime integers, we must have $\ell_1^{(a)} = \pm\ell_2^{(b)}$. The two D-branes then coincide over a set of n circles: the circle is the common closed line on the first torus, and n is the absolute value of the product of the intersection numbers on the second and third tori. If $\ell_1^{(a)}$ is displaced a little so that it does not coincide with $\ell_1^{(b)}$, then the two D-branes will also have no common points. With this understanding, we say that two branes with zero intersection number do not intersect.

21.2 D-branes and the Standard Model gauge group

We have seen that on the world-volume of N coincident D-branes there are $U(N)$ gauge fields, or gauge bosons, whose low energy dynamics is governed by a Yang–Mills theory

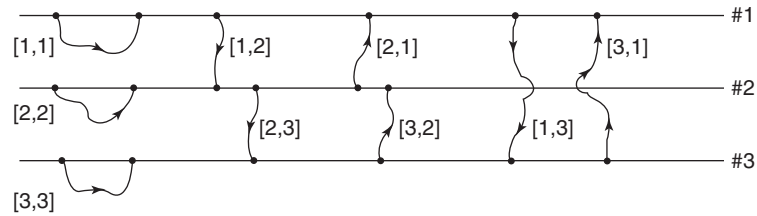

Fig. 21.4 **Three D-branes that give rise to the gauge bosons of $U(3)$. The branes are shown separate, for convenience, but are supposed to be coincident. We show nine strings, each belonging to one of the nine open string sectors.**

with gauge group $U(N)$. The gauge bosons of the Standard Model – the gluons, the $W^{\pm}$, the Z, and the photon – are all described by Yang–Mills theories. In this section we will examine the gauge group of the Standard Model, paying particular attention to the kind of D-brane configurations from which they could arise.

Let us start with the gluons, the gauge bosons that transmit the strong color force. Gluons are described by a four-dimensional $SU(3)$ Yang–Mills theory. This theory is closely related to the $U(3)$ Yang–Mills theory which arises at low energies on the world-volume of three coincident D3-branes (see Section 15.3). We will use the brane picture to understand what $SU(3)$ is. The D-brane configuration for $U(3)$ is shown in Figure 21.4, where the D-branes have been separated in order to be able to exhibit the various strings. There are nine open string sectors; they are labeled by $[ij]$, with $i, j = 1, 2, 3$. The figure shows one string from each sector. Every sector contains a string state that represents a gauge field.

Each of the three D-branes carries its own Maxwell field. These three Maxwell fields $A_{(i)\mu}$ $(i = 1, 2, 3)$ are associated with the states $\alpha_{-1}|[ii]\rangle$, which represent open strings that begin and end on the same D-brane. In writing $\alpha_{-1}|[ii]\rangle$, we have suppressed both the momentum labels of the states and the spacetime indices of the oscillators. With the help of these gauge fields, we can build a general class of states:

$$\sum_{i=1}^{3} A_{(i)}\, \alpha_{-1}|[ii]\rangle, \tag{21.10}$$

where we have suppressed the spacetime indices of the gauge fields, as well.

Let us now consider interactions, but restrict ourselves to low energies, where the string interactions become those of the Yang–Mills theory. The Maxwell fields $A_{(i)}$ do not interact with each other because the rule of combination of sectors (15.59) does not allow it: the endpoints of the different strings are never on the same brane. They do not have self-interactions, either, since Maxwell fields are free fields. The Maxwell fields $A_{(i)}$ do, however, interact with any state that carries their charge. Of the nine gauge fields in the D-brane configuration, the three Maxwell fields $A_{(i)}$ are on a special footing. Together, they comprise the largest possible set of gauge fields that have no interactions among its members.

Any string state is characterized by the values q_1, q_2, and q_3 of the charges that it carries with respect to the Maxwell fields $A_{(1)}$, $A_{(2)}$, and $A_{(3)}$, respectively. The charges of a state are collectively denoted by (q_1, q_2, q_3). In our convention, an oriented open string carries a unit negative charge at the $\sigma = 0$ endpoint and a unit positive charge at the $\sigma = \pi$ endpoint. A string in the [12] sector, for example, has charges $(-1, 1, 0)$. Here $q_1 = -1$ since the string begins on the first brane, $q_2 = +1$ since the string ends on the second brane, and $q_3 = 0$ since no endpoint lies on the third brane. For any string that is stretched from one brane to a different one, one charge is plus one, one charge is minus one, and one charge is zero. The sum of the three charges is zero. In fact, for a string beginning and ending on the same D-brane all three charges are zero.

If we think of a string in the limit in which it becomes a point with spacetime coordinates x^μ, the existence of the three charges implies the following coupling of the trajectory to the Maxwell fields:

$$\sum_{i=1}^{3} q_i \int A_{(i)} dx = \int \Big(\sum_{i=1}^{3} A_{(i)} q_i \Big) dx. \tag{21.11}$$

Because of (21.10), a change of basis in the space of states spanned by the $\alpha_{-1}|[ii]\rangle$ implies a linear redefinition of the Maxwell fields. On the other hand, (21.11) implies that a linear redefinition of the Maxwell fields results in a linear redefinition of the charges.

Since the idea of redefining fields and charges may be unfamiliar, let us give an example. Consider two Maxwell fields A_1 and A_2, and let (q_1, q_2) denote the charges of a given particle with respect to these Maxwell fields. The claim is that this physical system can be described using a different set of Maxwell fields, for which the charges are also different. To show this, we introduce new fields $A_\pm = (A_1 \pm A_2)/\sqrt{2}$ built from the original fields A_1 and A_2. In this setup, the sum on the right-hand side of (21.11) involves two terms only, and it can be rewritten as

$$\begin{aligned} A_1 q_1 + A_2 q_2 &= \frac{A_1 + A_2}{\sqrt{2}} \, \frac{q_1 + q_2}{\sqrt{2}} + \frac{A_1 - A_2}{\sqrt{2}} \, \frac{q_1 - q_2}{\sqrt{2}} \\ &= A_+ q_+ + A_- q_-, \end{aligned} \tag{21.12}$$

where we defined $q_\pm = (q_1 \pm q_2)/\sqrt{2}$. The physics, we claim, can be described using Maxwell fields A_+ and A_- and particles that carry charges $[q_+, q_-]$. Imagine, for example, two charged particles. In the first description (using A_1 and A_2), the first particle has charges $(1, 0)$, and the second particle has charges $(0, 1)$. It is clear that these particles do not experience an electrostatic force: the second particle carries no A_1 charge, and the first particle carries no A_2 charge. Now we describe the same two particles using fields A_+ and A_-. The new charges are $[1/\sqrt{2},\ 1/\sqrt{2}]$ for the first particle, and $[1/\sqrt{2},\ -1/\sqrt{2}]$ for the second particle. The force between the particles now has two contributions that cancel: the two particles carry the same A_+ charges, but they also carry A_- charges of the same magnitude but opposite signs. The net force is still zero. More generally, you can verify that the electrostatic force between arbitrarily charged particles does not change.

Quick calculation 21.5 Consider two particles with charges (q_1, q_2) and (q_1', q_2'), respectively. The electrostatic force between them is proportional to $q_1 q_1' + q_2 q_2'$. Let the

redefined charges of the two particles be $[q_+, q_-]$ and $[q'_+, q'_-]$, respectively. Show that $q_+q'_+ + q_-q'_- = q_1q'_1 + q_2q'_2$.

Returning to our brane configuration, we now claim that the nine gauge fields can be split into two sets, such that the gauge fields in one set have no interactions with the gauge fields in the other set. One set has eight gauge fields, and the other set has one gauge field. That special gauge field is the Maxwell field $\bar{A}_{(3)}$ associated with the state

$$|s_3\rangle \equiv \frac{1}{\sqrt{3}} \left(\alpha_{-1}|[11]\rangle + \alpha_{-1}|[22]\rangle + \alpha_{-1}|[33]\rangle\right). \tag{21.13}$$

Since Maxwell fields only interact with charged objects, it suffices to show that no gauge field carries its charge. In fact, we assert that the charge $\bar{q}_3$ associated with $\bar{A}_{(3)}$ is given by

$$\bar{q}_3 = \frac{1}{\sqrt{3}} \left(q_1 + q_2 + q_3\right). \tag{21.14}$$

If this is true, our claim holds; we have seen that the sum of the three charges is equal to zero for any string fully contained in the D-brane configuration.

To prove these claims, consider a change of basis states implemented by a general invertible linear transformation:

$$|s_i\rangle = \sum_{j=1}^{3} M_{ij}\, \alpha_{-1}|[jj]\rangle, \quad i = 1, 2, 3. \tag{21.15}$$

We can use matrix notation to write this equation as

$$\mathbf{s} = M\,\mathbf{v}. \tag{21.16}$$

Here $\mathbf{s}$ denotes the column vector with entries $|s_i\rangle$, $\mathbf{v}$ denotes the column vector with entries $\alpha_{-1}|[ii]\rangle$, and M is the invertible matrix with components M_{ij}. Letting $\mathbf{A}$ denote the row vector with entries $A_{(i)}$, we see that the general state in (21.10) can be rewritten as a dot product:

$$\mathbf{A}\cdot\mathbf{v} = \mathbf{A}\cdot M^{-1}M\mathbf{v} = \mathbf{A}M^{-1}\cdot\mathbf{s} \equiv \bar{\mathbf{A}}\cdot\mathbf{s}, \tag{21.17}$$

where $\bar{\mathbf{A}} = \mathbf{A}M^{-1}$ denotes the Maxwell fields associated with the new basis states. Equation (21.11) can now be used to find the new charges, which are the linear combinations of the old charges that multiply the new fields. Focusing on the pairing between fields and charges and letting $\mathbf{q}$ denote the column vector of charges q_i, we have

$$\sum_{i=1}^{3} A_{(i)}q_i = \mathbf{A}\cdot\mathbf{q} = \mathbf{A}M^{-1}\cdot M\mathbf{q} = \bar{\mathbf{A}}\cdot M\mathbf{q} \equiv \bar{\mathbf{A}}\cdot\bar{\mathbf{q}}. \tag{21.18}$$

Here the new charges $\bar{\mathbf{q}}$ are given in terms of the old charges $\mathbf{q}$ by

$$\bar{\mathbf{q}} = M\mathbf{q}. \tag{21.19}$$

We see that the matrix M that defines the new states (21.16) also defines the new charges. This is reasonable, since equations (21.10) and (21.11) show that states and charges couple to the fields in the same way. Our result proves that (21.13) implies (21.14). This is what we

had to establish in order to conclude that $\bar{A}_{(3)}$ decouples from the other eight gauge fields in the D-brane configuration. It can be proven that the other eight gauge fields cannot be split further into sets of fields that are mutually noninteracting.

There is one minor point that should also be addressed. In order to be able to read out charges correctly, the normalization of the F^2 terms in the action for the Maxwell fields must be preserved. Let $\mathbf{F}$ and $\bar{\mathbf{F}}$ denote the row vectors of field strengths obtained from the row vector gauge fields $\mathbf{A}$ and $\bar{\mathbf{A}}$, respectively. The relation $\mathbf{A} = \bar{\mathbf{A}}M$ then implies that $\mathbf{F} = \bar{\mathbf{F}}M$. The sum of F^2 terms is simply $\mathbf{F} \cdot \mathbf{F}$, where the entry to the right of the dot must be understood as a column vector, obtained by transposition from the row vector which is denoted with the same symbol. As a result, we have

$$\mathbf{F} \cdot \mathbf{F} = \bar{\mathrm{F}}M \cdot M^{\mathrm{T}}\bar{\mathrm{F}} = \bar{\mathrm{F}} \cdot MM^{\mathrm{T}}\bar{\mathrm{F}}. \tag{21.20}$$

In order to obtain $\bar{F}^2$ terms with the same normalization as that of the F^2 terms, we need $MM^{\mathrm{T}} = \mathbf{1}$ – the matrix M must be orthogonal. This is not a severe constraint; an orthogonal M is obtained if we supplement (21.13) with the states

$$\begin{aligned} |s_1\rangle &\equiv \frac{1}{\sqrt{6}}\left(\alpha_{-1}|[11]\rangle + \alpha_{-1}|[22]\rangle - 2\alpha_{-1}|[33]\rangle\right), \\ |s_2\rangle &\equiv \frac{1}{\sqrt{2}}\left(-\alpha_{-1}|[11]\rangle + \alpha_{-1}|[22]\rangle\right). \end{aligned} \tag{21.21}$$

Quick calculation 21.6 Write out the matrix M explicitly and verify that it is orthogonal.

It is interesting to note that the decoupled gauge field is obtained by adding up similar states from each of the three D-branes. We can also consider the state

$$\alpha^a_{-1}|[11]\rangle + \alpha^a_{-1}|[22]\rangle + \alpha^a_{-1}|[33]\rangle, \tag{21.22}$$

where the index a represents a direction normal to the D-branes. This state is associated with a displacement of the *full* collection of D-branes in the x^a direction (see the discussion below (15.37)). Just as the gauge field on a D-brane has massless scalar partners representing motion, the states above are partners of the decoupled gauge field.

We have mentioned before that the theory of a Maxwell gauge field is a $U(1)$ Yang–Mills theory. We have shown above that the $U(3)$ Yang–Mills theory of nine interacting gauge fields on three coincident D-branes contains a decoupled $U(1)$ theory, which we have identified explicitly. The remaining eight interacting gauge fields define the so-called $SU(3)$ gauge theory. This is the theory that governs the dynamics of the eight massless gluons of quantum chromodynamics (QCD). As groups, $U(3)$ is the group of 3-by-3 unitary matrices and $SU(3)$ is a subgroup of $U(3)$ obtained by considering only unitary matrices that have unit determinant. The relationship between the group $U(3)$ and the groups $SU(3)$ and $U(1)$ is written as

$$U(3) = SU(3) \times U(1), \tag{21.23}$$

where the product notation is used for groups that act independently. Equation (21.23) describes a local relation between groups which are also manifolds; it ignores topological

issues. Our physical analysis does not fully justify this notation, but it does show that the two factors correspond to gauge theories whose gauge fields do not interact with each other. The color force is described by $SU(3)$, so in order to obtain this theory from a configuration with three coincident D-branes we must understand the role of the additional Maxwell field. We will discuss this matter further in the following section. In general,

$$U(N) = SU(N) \times U(1), \tag{21.24}$$

where $U(N)$ is a theory with N^2 gauge fields and $SU(N)$ is a theory with $N^2 - 1$ gauge fields. As we have noted before, $U(N)$ gauge fields arise on the world-volume of N coincident D-branes.

The full set of Standard Model gauge bosons is described by the Yang–Mills theory with gauge group

$$SU(3)_c \times SU(2)_w \times U(1)_Y. \tag{21.25}$$

The subscript c is for color, w is for weak, and Y is for hypercharge. The $SU(2)_w \times U(1)_Y$ factors define the electroweak Yang–Mills theory. In $SU(2)_w$, there are $2^2 - 1 = 3$ gauge bosons. To realize $SU(2)_w$ with D-branes we need two additional coincident D-branes. This pair of D-branes should not coincide with the three color D-branes needed to obtain $SU(3)_c$, otherwise we would get a $U(5)$ Yang–Mills theory. If the two sets of branes are kept separate, the gauge group that emerges is

$$U(3) \times U(2) = SU(3) \times SU(2) \times U(1) \times U(1), \tag{21.26}$$

where we applied (21.24) twice and noted that the order of group factors has no significance. It is natural to ask if one of the $U(1)$ factors, or a combination of the two, can be identified with the hypercharge factor in (21.25). If we only cared about gauge bosons, the answer would be yes. In the Standard Model, however, we must also give the correct hypercharges to the fermions. This cannot be done with the $U(1)$s in (21.26). At least two additional D-branes appear to be necessary. We will discuss this matter further in the following section.

A process of symmetry breaking is necessary to reduce the gauge group in (21.25) to the one observed at low energies. When (21.25) applies, we get twelve massless gauge fields. At low energies, however, some of the Standard Model gauge fields are known to be massive. Indeed, three of the four gauge fields in $SU(2)_w \times U(1)_Y$ acquire mass through symmetry breaking, giving us the W^+, the W^-, and the Z^0. The fourth gauge field remains massless; it is the photon, which arises as a linear combination of the hypercharge field and one gauge field in the $SU(2)_w$ factor. After symmetry breaking, the gauge group is

$$SU(3)_c \times U(1)_{\text{em}}. \tag{21.27}$$

Symmetry breaking is triggered when certain charged scalar fields, the *Higgs* fields, acquire expectation values. In this process not only gauge fields, but also fermions, acquire mass. This brings us to the topic of fermions, which we discuss next.

21.3 Open strings and the Standard Model fermions

We have discussed the gauge bosons of the Standard Model, and we have begun to examine the D-brane configurations that are needed to obtain them. In order to use string theory to describe the full Standard Model, we must also learn about the matter particles and the charges that they carry. This is the main subject of the present section. We will also make preliminary observations regarding the representation of fermions as strings ending on the D-brane configurations that carry the gauge bosons. A detailed discussion of the embedding of the full Standard Model in a D-brane configuration is given in the following section.

Let us begin by examining some of the basic properties of fermions in four-dimensional spacetime. Consider for this purpose a prototype spin-$1/2$ fermion field; such a field represents a massive particle together with its (different) antiparticle. The classic example is the Dirac electron field, which describes the electron e^- and its antiparticle, the positron e^+. Imagine now a world where the mass of the electron and the mass of the positron are zero. The particle states would then be characterized by their *helicity*, the spin angular momentum along the direction of the motion. If the helicity takes the value $+1/2$ then the fermion state is said to be right handed. If the helicity takes the value $-1/2$ then the fermion state is said to be left handed. This is not altogether different from the way that we characterize photon states using a basis of right and left circularly polarized states; these are indeed photons of definite helicity. Massive fermion states can also be characterized by their helicity, but then the characterization is not Lorentz invariant. Imagine a massive fermion that moves with velocity $0 < v_x < c$ along the x axis of a Lorentz frame S and has spin angular momentum that points along the positive x axis. This is a fermion with positive helicity. In a Lorentz frame S' boosted along the x axis with velocity $v > v_x$, the particle moves in the negative x' direction, but its angular momentum does not reverse direction. In the S' frame the particle has negative helicity. If the particle is massless then it moves with the speed of light, and we cannot reverse its helicity by changing the Lorentz frame. We will be working with massless fermions.

Let f denote a fermionic particle and $\bar{f}$ denote the antiparticle. In this notation, e^- and e^+ correspond to f and $\bar{f}$, respectively. The quantum field theory which describes these particles will then include creation and annihilation operators for both the left- and right-handed states of the particle *and* for both the left- and right-handed states of the antiparticle. The creation operators, for example, would be written as:

$$(f_L^\dagger,\ f_R^\dagger),\quad (\bar{f}_L^\dagger,\ \bar{f}_R^\dagger), \tag{21.28}$$

where L and R stand for left-handed and right-handed, respectively. The operator $f_L^\dagger$, for example, creates a left-handed particle state when it acts on the vacuum, while $\bar{f}_R^\dagger$ creates a right-handed antiparticle state when it acts on the vacuum. The creation operators contain other labels (such as momentum) that we are suppressing for simplicity.

Quantum field theory automatically incorporates an important property of fermions: if we specify the charges of the *left*-handed *particles*, then the charges of the *right*-handed *antiparticles* are determined; they are in fact opposite. If the charges are specified as electric charges for a set of noninteracting Maxwell fields, then opposite charges are simply charges of opposite sign. Similarly, the charges of the right-handed particles and those of the left-handed antiparticles are also opposite. To display these relationships, we write

$$\begin{aligned} f_L^\dagger &\leftarrow \text{opposite charge} \rightarrow \bar{f}_R^\dagger, \\ f_R^\dagger &\leftarrow \text{opposite charge} \rightarrow \bar{f}_L^\dagger, \end{aligned} \tag{21.29}$$

where we have omitted, for brevity, the vacuum states. All charges are determined if we specify the charges of two states that do not appear on the same line. These charges can be specified independently. It suffices, for example, to fix the charges of the particles $f_L^\dagger$ and $f_R^\dagger$ to determine the charges of the antiparticles $\bar{f}_L^\dagger$ and $\bar{f}_R^\dagger$, or vice versa. Alternatively, we can fix the charges of the left-handed states $f_L^\dagger$ and $\bar{f}_L^\dagger$, or the charges of the right-handed states $f_R^\dagger$ and $\bar{f}_R^\dagger$. We will generally specify fermion charges by listing those of the left-handed states.

A central property of the Standard Model is that the spectrum of fermions is *chiral*. To understand this property we first consider a simpler case. Assume we have a theory whose entire fermion spectrum consists of the states in (21.29). The fermion is said to be chiral if the left- and right-handed particle states $f_L^\dagger$ and $f_R^\dagger$ do not have the same charges:

$$\text{if } f_L^\dagger \text{ and } f_R^\dagger \text{ do not have the same charges, the fermion is chiral.} \tag{21.30}$$

In fact, quantum field theory does not require both lines of (21.29) to exist in a theory. A theory can be consistent with only the states on the first line, or only the states on the second line. In such cases, unless the states are neutral, the fermion is automatically said to be chiral.

Charge describes the response of a particle to gauge bosons, so the left- and right-handed particle states of a chiral fermion respond differently to the same set of gauge bosons. Since the particle and antiparticle charges are correlated (see (21.29)), the left- and right-handed antiparticle states also respond differently to the same set of gauge bosons. The electron is chiral: left-handed electrons and right-handed electrons respond differently to the weak interactions, and so do left- and right-handed positrons. In fact, all the fermions in the standard model are chiral. Chirality is a very powerful property: in a gauge theory with chiral fermions, the fermions cannot acquire mass so long as the gauge symmetry that acts

chirally remains unbroken. Thus the term chiral fermions is synonymous with massless fermions. In the Standard Model, the electroweak interactions $SU(2)_w \times U(1)_Y$ act chirally. The Standard Model fermions remain massless until symmetry breaking reduces the gauge group (21.25) down to $SU(3)_c \times U(1)_{\text{em}}$. Since neither the color force nor the electromagnetic force act chirally, the fermions can then acquire mass. The mass scale is set by the mass parameter that appears in the symmetry breaking sector of the theory. This sector is called the Higgs sector. As we mentioned earlier, the Higgs bosons are charged scalar fields that trigger symmetry breaking when they acquire expectation values.

Quick calculation 21.7 Convince yourself that a fermion is chiral if the charges of the left-handed particles and those of the left-handed antiparticles are *not* opposite.

Let us now describe the matter content of the Standard Model. The fermions of the Standard Model fall into three *generations*, or sets, all of which contain the same number of particles with exactly the same charges. Once the fermions acquire masses, the three generations are no longer identical. There is a hierarchy of masses; the first generation contains the lightest fermions and the third generation contains the heaviest fermions. Within each generation we have quarks and leptons. Quarks feel both the strong interactions and the electroweak interactions, while leptons feel only the electroweak interactions. Our discussion will focus on charges, so it suffices to consider a single generation. We will give the charges of all the left-handed states in one generation of quarks and leptons. In other words, we will list the charges of all the left-handed particles and left-handed antiparticles.

Consider first the quarks. The quarks feel the $SU(3)$ color force because they carry color. Quarks come in three colors, which we will call red (r), blue (b), and green (g). The left-handed states q_L of a quark are therefore of three types:

$$q_{L\text{r}},\ q_{L\text{b}}, \quad \text{and} \quad q_{L\text{g}}. \tag{21.31}$$

The collection of these three states is said to form the representation **3** of the group $SU(3)$. One writes $q_L \sim \mathbf{3}$. Under an $SU(3)$ gauge transformation, the three states are rotated by an $SU(3)$ matrix. Since the color force does not act chirally, the left-handed antiquarks $\bar{q}_L$ carry the opposite color charges. These are called anti-red (a-r), anti-blue (a-b), and anti-green (a-g):

$$\bar{q}_{L\text{a-r}},\ \bar{q}_{L\text{a-b}}, \quad \text{and} \quad \bar{q}_{L\text{a-g}}. \tag{21.32}$$

One writes $\bar{q}_L \sim \overline{\mathbf{3}}$, the bar included to mean opposite. In group theory, the **3** and the $\overline{\mathbf{3}}$ are said to be conjugate representations. When the states (21.31) are acted upon by an $SU(3)$ matrix M, the states (21.32) are acted upon by the complex conjugate matrix M^*.

This discussion about quarks and $SU(3)$ representations can be made more intuitive using D-branes and the open strings that can represent the quarks. We have seen that $SU(3)_c$ requires three coincident D-branes. The key insight is that quarks are simply open strings that have *one* endpoint on one of these three branes (recall that gluons have *both* endpoints on the collection of branes). We can use the labels red, blue, and green to refer to the first, second, and third brane, respectively. An open string that ends on the red brane is a left-handed red quark, a string that ends on the blue brane is a left-handed blue quark, and a

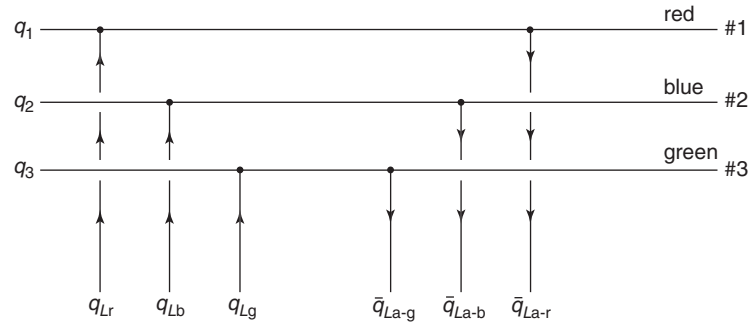

Fig. 21.5 **The left-handed red, blue, and green quarks are open strings that *end* on the color D-branes. The left-handed anti-red, anti-blue, and anti-green antiquarks are open strings that *begin* on the color D-branes.**

string that ends on the green brane is a left-handed green quark. The orientation of these strings points into the branes, or, equivalently, the endpoints on the brane are the $\sigma = \pi$ endpoints. How do the left-handed antiquarks arise? They are simply oppositely oriented open strings. An open string starting on the red brane, for example, would be a left-handed anti-red antiquark. All these strings are shown in Figure 21.5. We will refer to the three branes of $SU(3)_c$ as *color* branes, or *baryonic* branes. The name baryonic comes from baryon, a particle that is composed of three quarks.

We can also describe the $SU(3)$ charges of the quarks as follows. Recall that we characterize string states by their charges (q_1, q_2, q_3) with respect to the Maxwell fields that live on the branes. These three charges in fact define the $U(3)$ charge of a state. The charge for the decoupled $U(1)$ is proportional to $(q_1 + q_2 + q_3)$. We can label the $SU(3)$ charges using any *two* linear combinations of the charges that, together with $(q_1 + q_2 + q_3)$, define a linearly independent set. It is convenient to employ the pair (a_1, a_2) defined by

$$(a_1, a_2) \equiv (\, q_1 - q_2,\ q_2 - q_3\,). \tag{21.33}$$

The three left-handed quarks in (21.31), which comprise the representation **3** of $SU(3)$, are therefore characterized by

$$\mathbf{3}\colon \quad (1, 0),\ \ (-1, 1),\ \ (0, -1). \tag{21.34}$$

For the three left-handed antiquarks in (21.32) we have

$$\overline{\mathbf{3}}\colon \quad (-1, 0),\ \ (1, -1),\ \ (0, 1). \tag{21.35}$$

In the language of representation theory, a pair (a_1, a_2) is called a weight vector, and the entries a_1 and a_2 are the Dynkin labels of the weight vector. The **3** and the $\overline{\mathbf{3}}$ representations have three weight vectors.

Quick calculation 21.8 Give the eight weight vectors corresponding to the gauge field states in $SU(3)$. These states define the representation **8**, and the weights are the gluon

charges. Each of the three gluons that begin and end on the same brane has weight $(0, 0)$ (why?). Only two of these belong to **8**, since one of them corresponds to the decoupled $U(1)$.

Only one endpoint of any open string that represents a quark lies on a color D-brane, so it is natural to ask: where does the other endpoint lie? For the left-handed quarks, we can find the answer by looking into their $SU(2)_w$ charges. Quark states fall into $SU(2)$ representations. The representations of $SU(2)$ are familiar from their role in the quantum mechanics of spin angular momentum. The $SU(2)_w$ of weak interactions has no relation to angular momentum, so one speaks of *isospin* instead of spin. The representation of isospin $I = 1/2$, for example, has two states, one with $I_3 = 1/2$ and the other with $I_3 = -1/2$, where I_3 denotes a third component of isospin.

Left-handed quark states fall into $I = 1/2$ representations of the weak interactions. New labels, called flavor labels, are needed to characterize the various types of quarks. For any fixed color, the left-handed u-quark is a state with $I_3 = 1/2$, and the (same color) left-handed d-quark is a state with $I_3 = -1/2$. Appropriately, the u and d stand for up and down, respectively. Thus u_L and d_L are members of an $SU(2)_w$ doublet, which we denote by **2**. Since quarks come in three colors, there are three $SU(2)$ doublets. Let us now think in terms of D-branes. We can build a $U(2) = SU(2)_w \times U(1)$ theory with two coincident D-branes. The left-handed u-quarks are strings that begin on one of these D-branes, call it the first brane, and end on one of the color branes. The left-handed d-quarks are strings that begin on the other D-brane, call it the second brane, and end on one of the color branes. The two D-branes we have introduced are called *left* branes. Left-handed quarks are open strings that begin on a left brane and end on a color brane.

The states of $SU(2)$ representations which arise from strings that have an endpoint on the left branes can also be characterized by charges. This time we have two charges $\bar{q}_1$ and $\bar{q}_2$ associated with the Maxwell fields living on brane one and brane two, respectively. As usual, in our convention $\bar{q}_i = +1$ for a string that ends on the brane i, and $\bar{q}_i = -1$ for a string that begins on the brane i. The charges $\bar{q}_1$ and $\bar{q}_2$ define the $U(2)$ charge of a state. The charge for the decoupled $U(1)$ is proportional to $(\bar{q}_1 + \bar{q}_2)$. We can label the $SU(2)$ charge using any combination of the charges that is independent from $(\bar{q}_1 + \bar{q}_2)$. It is convenient to use a Dynkin label

$$a_1 \equiv \bar{q}_1 - \bar{q}_2. \tag{21.36}$$

A string that begins on brane one has charges $(-1, 0)$, so $a_1 = -1$. Since this is a u-quark, which has $I_3 = 1/2$, we deduce a linear relation between the third component of isospin and the Dynkin label a_1:

$$I_3 = -a_1/2. \tag{21.37}$$

This relation also works for the d-quark: a string that begins on brane two has charges $(0, -1)$, which gives $a_1 = 1$ and $I_3 = -1/2$. Equation (21.37) illustrates how the quantum numbers of the states in a representation are related to the charges of the states.

Let us now try to visualize the full configuration of three coincident color branes and two coincident left branes. Whether or not the two sets of D-branes coincide, as long as they

are parallel, the quark states stretching from one group of branes into the other fail to correspond to the Standard Model quarks in a dramatic way: they are not chiral. It can be shown that the spectrum contains left-handed quarks *and* right-handed quarks, with exactly the same charges. In a configuration with parallel D-branes, all states in (21.29) are in fact produced, and the left-handed and right-handed particles carry the same charges. The lack of chirality is manifest when the sets of D-branes are parallel and separate: the stretched strings are massive, and masses are not allowed for chiral fermions.

A physical situation is obtained if the coincident color D-branes *intersect* the coincident left D-branes. What we have in mind here is a configuration of intersecting D-branes of the type considered in Section 21.1. If the two sets of D-branes intersect, the fermion fields represented by strings that stretch from one set to the other will be localized near the intersection. Imagine beginning with coincident and parallel branes, where open strings can move in all the spatial directions along the branes. As the intersection angle grows away from zero, it is as if compactification had occurred: some directions of motion are lost. Some states that are massless at zero angle become massive. In fact, under suitable conditions, only the states in *one* line of (21.29) remain massless. We get a chiral fermion precisely because *only one* of the pairs is produced at the intersection: either a left-handed particle and its partner, or a left-handed antiparticle and its partner. If one pair is produced, the other pair may or may not be produced elsewhere on the D-brane configuration. There will be no reason for their charges to be correlated, if it is produced. This is why brane-intersection models naturally give a chiral fermion spectrum. In order to decide if a fermion spectrum is actually chiral, we must find all the fermion states that arise from the complete brane configuration.

We can now use Figure 21.6 to give a partial description of the situation. The color branes are shown horizontally, and the left branes are shown vertically. The three left-handed u-quarks and the three left-handed d-quarks are shown as stretched strings. The oppositely oriented strings (not shown) correspond to the oppositely charged right-handed antiquarks. The I_3 labels on the left branes indicate the values of I_3 for a string that *ends* on the branes.

The specification of charges for the left-handed quarks is completed by stating the values of the hypercharge Y. All three u_L quarks and all three d_L quarks are states of $Y = 1/6$. The information about representations and charges of fermions in the Standard Model is usually summarized using the notation

$$\big(\text{color},\ \text{isospin}\big)_Y. \tag{21.38}$$

Here color and isospin stand for the representations of $SU(3)$ and $SU(2)$, respectively. Y is simply the value of the hypercharge. This notation describes representations of the full Standard Model gauge group $SU(3) \times SU(2) \times U(1)$, since it specifies the representations with respect to each of the group factors. To illustrate the use of this notation, consider the states

$$\begin{pmatrix} u_{L\mathrm{r}} & u_{L\mathrm{b}} & u_{L\mathrm{g}} \\ d_{L\mathrm{r}} & d_{L\mathrm{b}} & d_{L\mathrm{g}} \end{pmatrix}, \tag{21.39}$$

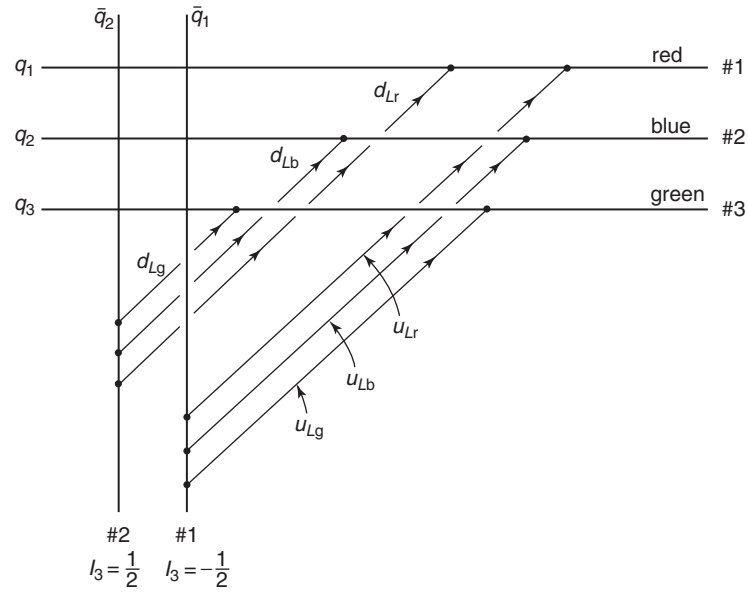

Fig. 21.6 **The left-handed quarks are strings that stretch from the left branes to the baryonic branes. The three left-handed *u*-quarks are open strings that begin on the first left brane and end on a baryonic brane. The three left-handed *d*-quarks are open strings that begin on the second left brane and end on a baryonic brane.**

which comprise the left-handed up and down quarks of all possible colors. The states can be viewed as a triplet of $SU(2)$ doublets (the three columns of the matrix) or, equivalently, as a doublet of $SU(3)$ triplets (the two rows of the matrix). In the notation of (21.38), the six states are denoted by

$$(\mathbf{3}, \mathbf{2})_{1/6}\,. \tag{21.40}$$

Since the labels for the color and isospin representations are actually equal to the number of states in the representations, the product $3 \times 2 = 6$ is the number of states in the full representation. All six states have the same hypercharge.

How do we calculate the hypercharge of states from the D-brane configuration? It turns out that the hypercharge receives contributions from both the decoupled $U(1)$ of the baryonic branes and the decoupled $U(1)$ of the left branes. Let us define conveniently normalized $U(1)$ charges Q_1 and Q_2, associated with the color and left branes, respectively:

$$Q_1 = q_1 + q_2 + q_3, \quad Q_2 = \bar{q}_1 + \bar{q}_2. \tag{21.41}$$

Any string that begins and ends on baryonic branes or that begins and ends on left branes has both Q_1 and Q_2 equal to zero. On the other hand, a left-handed quark is a string that begins on a left brane and ends on a color brane. Any such string has $Q_1 = 1$ and $Q_2 = -1$. The hypercharge of the left-handed quarks is obtained if we take

$$Y = -\tfrac{1}{3}\,Q_1 - \tfrac{1}{2}\,Q_2 - \cdots, \tag{21.42}$$

where the dots represent contributions from additional D-branes, to be included later. Since the left-handed quarks have no endpoints on these extra branes, $Y = -(1/3) \times 1 - (1/2) \times (-1) = 1/6$.

We must now consider the left-handed *antiquarks*, $\bar{u}_L$ and $\bar{d}_L$. At this point, we encounter the chirality of the electroweak interactions. Had the electroweak interactions been non-chiral, these antiquarks would form a doublet with charges opposite to those of the left-handed u and d quarks; they would be seen as strings ending on the left branes. It turns out, however, that the left-handed *antiquarks* $\bar{u}_L$ and $\bar{d}_L$ are each $SU(2)$ singlets (with representation denoted by **1**). The corresponding strings cannot then have endpoints on the left branes. Moreover, $\bar{u}_L$ has $Y = -2/3$ and $\bar{d}_L$ has $Y = 1/3$. A nonchiral coupling would have required $Y = -1/6$ for both of them. It follows from this and our earlier analysis of color that the left-handed antiquarks are in the representations

$$\bar{u}_L \sim \left(\bar{\mathbf{3}}, \mathbf{1}\right)_{-2/3} \quad \text{and} \quad \bar{d}_L \sim \left(\bar{\mathbf{3}}, \mathbf{1}\right)_{1/3}. \tag{21.43}$$

Here we have suppressed the color labels; thus $\bar{u}_L$, for example, stands for the three antiquarks of different colors. The representations of the left-handed quarks and antiquarks are then summarized as

$$\begin{pmatrix} u_L \\ d_L \end{pmatrix} \sim (\mathbf{3}, \mathbf{2})_{1/6}\,, \quad \bar{u}_L \sim \left(\bar{\mathbf{3}}, \mathbf{1}\right)_{-2/3} \quad \text{and} \quad \bar{d}_L \sim \left(\bar{\mathbf{3}}, \mathbf{1}\right)_{1/3}. \tag{21.44}$$

This is the set of left-handed quark and antiquark states in the first generation of the Standard Model. The generation also includes the corresponding right-handed states.

The $\bar{u}_L$ antiquarks are strings that begin on a color brane. Since they cannot end on a left brane (they are $SU(2)_w$ singlets), they must end on a new D-brane. If we let Q_3 denote the electric charge that couples to the Maxwell field on this D-brane, we find that (21.42) must be changed to

$$Y = -\tfrac{1}{3}\, Q_1 - \tfrac{1}{2}\, Q_2 - Q_3 - \cdots. \tag{21.45}$$

The hypercharge $Y = -2/3$ of the $\bar{u}_L$ states emerges because $Q_1 = -1$, $Q_2 = 0$, and $Q_3 = 1$. The $\bar{d}_L$ antiquarks are also strings that begin on a color brane and cannot end on a left brane. Their hypercharge $Y = 1/3$ is correctly reproduced by (21.45) with $Q_1 = -1$ and $Q_2 = 0$, if the strings end on a D-brane that does not contribute to Y. We will discuss such additional D-branes in the next section.

Let us now describe the left-handed leptons in the first generation. These are the left-handed electron-neutrino and the left-handed electron, together with the left-handed anti-neutrino and the left-handed positron. None of the leptons carries color; they are all color singlets, and their $SU(3)$ representation is denoted by **1**. The left-handed neutrinos and the left-handed electrons form an $SU(2)$ doublet with hypercharge $Y = -1/2$. Again, we have chirality. Both the left-handed positron and the left-handed anti-neutrino are $SU(2)$

singlets, and their hypercharges are one and zero, respectively. The lepton charges are summarized by

$$\begin{pmatrix} \nu_{eL} \\ e_L^- \end{pmatrix} \sim (\mathbf{1},\mathbf{2})_{-1/2}\,, \quad e_L^+ \sim (\mathbf{1},\mathbf{1})_1\,, \quad \text{and} \quad \bar{\nu}_{eL} \sim (\mathbf{1},\mathbf{1})_0\,. \tag{21.46}$$

Note that the left-handed anti-neutrinos are color, weak, and hypercharge singlets. These states have not been detected directly, but they are likely to exist because neutrinos appear to have mass. The states in (21.44) and (21.46), together with the corresponding right-handed states, comprise the matter states in the first family of the Standard Model. As you have probably already realized, the open strings that represent leptons cannot end on color D-branes. The left-handed neutrino and the left-handed electron arise from strings that have one endpoint on a left brane and another endpoint on some other brane. Neither the left-handed positron nor the left-handed anti-neutrino have an endpoint on a color brane or a left brane.

The full set of left-handed states in a generation is obtained by listing the states in (21.44) together with those in (21.46). Suppressing the names of the states, and using the + symbol to collect the representations together, we have

$$(\mathbf{3},\mathbf{2})_{1/6} + \left(\overline{\mathbf{3}},\mathbf{1}\right)_{-2/3} + \left(\overline{\mathbf{3}},\mathbf{1}\right)_{1/3} + (\mathbf{1},\mathbf{2})_{-1/2} + (\mathbf{1},\mathbf{1})_1 + (\mathbf{1},\mathbf{1})_0\,. \tag{21.47}$$

As we have learned, charge reversal exchanges the **3** and $\overline{\mathbf{3}}$ of $SU(3)$ and reverses the sign of the hypercharge. Under charge reversal all singlets **1** remain unchanged, because they represent states with zero charge. For $SU(2)$, charge reversal exchanges the **2** and $\overline{\mathbf{2}}$, but it turns out that the $\overline{\mathbf{2}}$ is a representation that is equivalent to the **2**. So charge reversal leaves the **2** of $SU(2)$ unchanged.

On account of the comments below (21.29), we can list the states in a generation in alternative ways.

Quick calculation 21.9 List the right-handed states in a generation, together with their corresponding charges.

Quick calculation 21.10 Describe the matter states in a generation by listing the left- and right-handed *particle* states, together with their corresponding charges.

In the Standard Model, the electric charge Q_{em} arises from a linear combination of the hypercharge Y and the third component I_3 of isospin:

$$Q_{\text{em}} = Y + I_3. \tag{21.48}$$

If we apply this formula to the u_L and d_L states, we find

$$\begin{aligned} Q_{\text{em}}(u_L) &= Y(u_L) + I_3(u_L) = \tfrac{1}{6} + \tfrac{1}{2} = +\tfrac{2}{3}, \\ Q_{\text{em}}(d_L) &= Y(d_L) + I_3(d_L) = \tfrac{1}{6} - \tfrac{1}{2} = -\tfrac{1}{3}. \end{aligned} \tag{21.49}$$

These are indeed the correct values. The proton is made up of two up quarks and one down quark. Its electric charge equals $2 \times \frac{2}{3} - \frac{1}{3} = 1$.

Quick calculation 21.11 Find the electric charges of the left-handed antiquarks, and verify that they are opposite to those of the left-handed quarks. Show that (21.48) gives zero electric charge to the left-handed neutrino and anti-neutrino states, electric charge minus one to the left-handed electron states, and electric charge plus one to the left-handed positron states. Conclude that electromagnetism does not couple chirally.

As far as states and charges are concerned, the other two generations of the Standard Model are copies of the first generation. The quark flavors in the second generation are called charm and strange, and denoted by c and s respectively. The leptons in the second generation are the muon-neutrino ν_μ and the muon μ^-. In the third generation we have a top quark t and a bottom quark b, as well as a tau-neutrino ν_τ and a tau τ^-. The full set of left-handed states in the Standard Model consists of three copies of the states in (21.47):

$$3 \times \left[(\mathbf{3}, \mathbf{2})_{1/6} + \left(\overline{\mathbf{3}}, \mathbf{1}\right)_{-2/3} + \left(\overline{\mathbf{3}}, \mathbf{1}\right)_{1/3} + (\mathbf{1}, \mathbf{2})_{-1/2} + (\mathbf{1}, \mathbf{1})_1 + (\mathbf{1}, \mathbf{1})_0 \right]. \tag{21.50}$$

We can now state precisely the chirality property of the fermion spectrum in the Standard Model. A spectrum is said to be *non*chiral if the set of left-handed states can be split into pairs of left-handed states with opposite charges (see Quick calculation 21.7). Thus, given a list of charged left-handed states that defines a nonchiral spectrum, the operation of reversing the charges of all the states must leave the *list* of states invariant. The fermion spectrum of the standard model is chiral because the operation of reversing all charges in the list (21.50) changes the list.

The gauge group and the matter content of the Standard Model may seem to you rather intricate or perhaps even cumbersome. But this set of particles and interactions in fact provides a rather economical description of an extremely large number of experimental results obtained over the past few decades. The Standard Model of particle physics is indeed a magnificent achievement. It is not a final theory of particle physics, nor is it a complete one, but it seems certain that the Standard Model must appear in the low energy limit of any correct unified theory of all interactions. It is in this sense that the Standard Model of particle physics has become a permanent part of our knowledge about the physical world.

21.4 The Standard Model on intersecting D6-branes

In the two previous sections we acquainted ourselves with some of the ingredients needed to build a string theory model of elementary particles. In this section, we build a complete string model that has many of the features of the Standard Model. The model involves intersecting D6-branes wrapped on a T^6, in the framework of type IIA superstring theory.

Before symmetry breaking, it has all the massless particles of the Standard Model, but it also contains a few extra particles. It is in fact possible to construct a model with orientifold O6-planes and D6-branes that gives *precisely* the particle content of the Standard Model. An orientifold Op-plane is an extended object with p spatial dimensions. The basic properties of orientifold planes were studied earlier in a series of problems (13.6, 15.1, 15.3, 15.4, and 15.5). We will introduce some of the features of this second model, relegating most of the analysis to the problems at the end of the chapter.

One key property of the Standard Model matter content is its replication: there are three generations that contain fermionic matter states with identical charges. How do we explain such replication in terms of D-branes? We explain it in terms of multiple intersections. We found the left-handed u and d quarks at the intersection of the coincident baryonic branes with the coincident left branes. If these sets of branes wrap around the T^6 in such a way that they intersect three times, then the second intersection will give left-handed c and s quarks, and the third intersection will give left-handed t and b quarks. We learned in Section 21.1 that general classes of D6-branes wrapped on $T^6 = T^2 \times T^2 \times T^2$ are characterized by three lines ℓ_1, ℓ_2, and ℓ_3, each of which is, in turn, characterized by two integers. Moreover, we found the formula (21.9), which gives the number of times that two D6-branes intersect.

Let $N_1 = 3$ denote the number of baryonic branes, and let $N_2 = 2$ denote the number of left branes. Moreover, let them be wrapped on T^6 as follows:

$$\begin{aligned} N_1 = 3: &\quad \ell_1^{(1)} = (1,2), \quad \ell_2^{(1)} = (1,-1), \quad \ell_3^{(1)} = (1,-2), \\ N_2 = 2: &\quad \ell_1^{(2)} = (1,1), \quad \ell_2^{(2)} = (1,-2), \quad \ell_3^{(2)} = (-1,5). \end{aligned} \tag{21.51}$$

The intersection number I_{12} between a single baryonic brane and a single left brane is then

$$I_{12} = \big(1\times 1 - 1\times 2\big)\cdot\big(1\times(-2) - 1\times(-1)\big)\cdot\big(1\times 5 - (-1)\times(-2)\big) = 3. \tag{21.52}$$

The intersection number is the desired one. Of course, the constraint $I_{12} = 3$ does not determine the wrappings in (21.51). These wrapping numbers are one of many possibilities that work.

We found that the left-handed quarks are obtained as strings that stretch from the N_2 branes to the N_1 branes. The orientation of these strings is correlated with the sign of I_{ab}. Let I_{ab} denote the intersection number for a D6-brane a and a D6-brane b. The precise rule that gives the number of fermions and specifies how they are represented by strings is given as follows.

> There are $|I_{ab}|$ left-handed fermions at the intersection set of brane a and brane b, one left-handed fermion at each intersection point. If $I_{ab} > 0$, the left-handed states are strings that stretch from brane b to brane a. If $I_{ab} < 0$, the left-handed states are strings that stretch from brane a to brane b.

The orientation of the strings determines the charges of the left-handed states. The states produced at an intersection are chiral in the sense discussed below (21.30): only the minimal set of states is produced. The oppositely oriented strings at the intersection represent the oppositely charged right-handed *antiparticles* that must necessarily accompany the left-handed particle states.

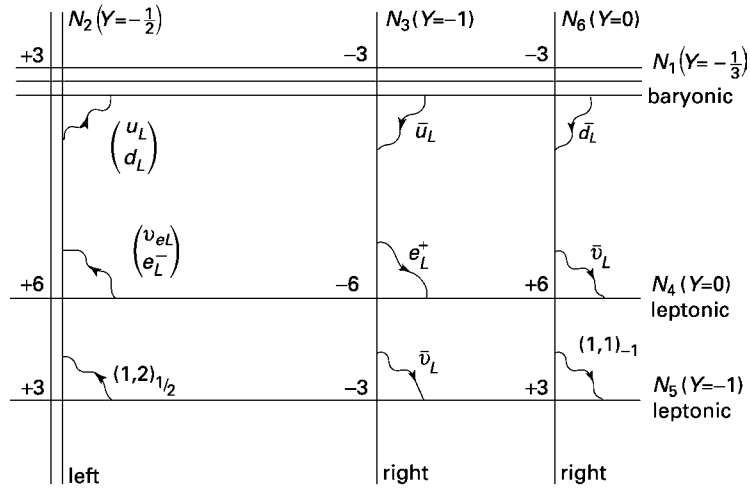

Fig. 21.7 **The brane configuration consisting of $N_1 = 3$ baryonic branes, $N_2 = 2$ left branes, $N_3 = N_6 = 1$ right branes, and $N_4 = N_5 = 1$ leptonic branes. The values of Y indicate the contribution to the hypercharge for a string that ends on the brane. At the intersections we give the intersection numbers I_{ab} with $a < b$. The name right branes is used because right-handed *particles* are attached to them (we actually show the left-handed antiparticles).**

For the three generations of left-handed quarks obtained above, the value $Y = 1/6$ of the hypercharge is correctly given by (21.42). The picture of the branes is that of Figure 21.6, with the understanding that the displayed intersection actually occurs two additional times. To indicate this on the diagram, we insert a $+3$ at the intersection, as shown in Figure 21.7. We now try to obtain the three generations of left-handed antiquarks. Consider first the left-handed u-antiquark and its two copies (the left-handed c and t antiquarks). The open string that represents this quark begins on a baryonic brane, but it cannot end on a left brane because the state is an $SU(2)$ singlet. We thus need a new D-brane. Let $N_3 = 1$ represent another D-brane. According to the rule above, we need $I_{13} = -3$, since the open strings must start on the baryonic branes. Additionally, the contribution to the hypercharge from the charge Q_3 on this brane must be such that $Y(\bar{u}_L) = -2/3$. This requires that (21.42) be replaced by

$$Y = -\tfrac{1}{3}\,Q_1 - \tfrac{1}{2}\,Q_2 - Q_3 - \cdots. \tag{21.53}$$

We then find $Y(\bar{u}_L) = -(1/3)(-1) - (1/2)(0) - 1 = -2/3$, as desired. We fix the wrapping of the brane on T^6 by

$$N_3 = 1: \quad \ell_1^{(3)} = (1,1), \quad \ell_2^{(3)} = (1,0), \quad \ell_3^{(3)} = (-1,5). \tag{21.54}$$

Including the result in (21.52), we then find

$$I_{12} = 3, \quad I_{13} = -3, \quad I_{23} = 0. \tag{21.55}$$

We see that the N_3 brane does not intersect the left branes. The N_3 brane is shown in Figure 21.7 as a vertical brane, placed to the right of the left branes. The equation $Y = -1$

is added to the label of the brane to indicate that a string that ends on this brane receives a contribution to the hypercharge equal to minus one (this number is equal to the coefficient of Q_3 in (21.53)). Similar labels are attached to the baryonic and to the left branes. The N_3 brane is called a right brane because right-handed quark states are attached to it. Indeed, instead of the $\bar{u}_L$ antiquark shown in the figure, we could have exhibited the oppositely oriented string which gives the right-handed quark state u_R.

In this model we still have to add three additional D-branes: another right brane to obtain the left-handed $\bar{d}_L$ antiquarks (and its copies) and two separate *leptonic* branes to obtain the leptons. Since we are not in a position to derive the wrappings of the D-branes, let us simply state them. Together with the

$$\begin{aligned} N_4 = 1: &\quad \ell_1^{(4)} = (1,2), \quad \ell_2^{(4)} = (-1,1), \quad \ell_3^{(4)} = (1,1), \\ N_5 = 1: &\quad \ell_1^{(5)} = (1,2), \quad \ell_2^{(5)} = (-1,1), \quad \ell_3^{(5)} = (2,-7), \\ N_6 = 1: &\quad \ell_1^{(6)} = (1,1), \quad \ell_2^{(6)} = (3,-4), \quad \ell_3^{(6)} = (1,-5). \end{aligned} \tag{21.56}$$

Both the N_4 and the N_6 branes are declared not to contribute to the hypercharge, but the N_5 brane is declared to contribute minus one unit. So the final formula for the hypercharge reads

$$Y = -\tfrac{1}{3}\, Q_1 - \tfrac{1}{2}\, Q_2 - Q_3 - Q_5. \tag{21.57}$$

The remaining intersection numbers are easily calculated and the complete list is

$$\begin{aligned} &I_{12} = 3, \quad I_{13} = -3, \quad I_{23} = 0, \\ &I_{14} = 0, \quad I_{15} = 0, \quad I_{16} = -3, \\ &I_{24} = 6, \quad I_{25} = 3, \quad I_{26} = 0, \\ &I_{34} = -6, \quad I_{35} = -3, \quad I_{36} = 0, \\ &I_{45} = 0, \quad I_{46} = 6, \quad I_{56} = 3. \end{aligned} \tag{21.58}$$

Since $I_{ab} = -I_{ba}$, we have only listed intersection numbers I_{ab} with $a < b$. Those are the values exhibited at the intersections in Figure 21.7.

Quick calculation 21.12 Confirm the values of all the intersection numbers in (21.4), and check that they are correctly recorded in the figure. Check also that the strings shown in the figure have the orientation required by the rule for left-handed states.

We can organize the list of intersections using the horizontal branes on the figure, all of which intersect with the two coincident left branes and the two right branes. The intersections on the baryonic branes give

$$\underbrace{3\,(\mathbf{3},\mathbf{2})_{1/6}}_{[12]} + \underbrace{3\left(\bar{\mathbf{3}},\mathbf{1}\right)_{-2/3}}_{[13]} + \underbrace{3\left(\bar{\mathbf{3}},\mathbf{1}\right)_{1/3}}_{[16]}, \tag{21.59}$$

where the $[ab]$ labels under the representations indicate that the corresponding fermions arise from the I_{ab} intersection. The above representations are precisely three copies of a

single generation of quarks, as you can see by comparing with (21.44). The intersections on the leptonic brane N_4 give

$$\underbrace{6\,(\mathbf{1},\mathbf{2})_{-1/2}}_{[24]} + \underbrace{6\,(\mathbf{1},\mathbf{1})_{1}}_{[34]} + \underbrace{6\,(\mathbf{1},\mathbf{1})_{0}}_{[46]}\,. \tag{21.60}$$

This result justifies the name leptonic. The above representations are six copies of a single generation of leptons, as you can see by comparing with (21.46). This is more than the three copies that we needed, but there is even more. The intersections on the leptonic N_5 brane give us

$$\underbrace{3\,(\mathbf{1},\mathbf{2})_{1/2}}_{[25]} + \underbrace{3\,(\mathbf{1},\mathbf{1})_{0}}_{[35]} + \underbrace{3\,(\mathbf{1},\mathbf{1})_{-1}}_{[56]}\,. \tag{21.61}$$

Quick calculation 21.13 Use (21.57) and the picture of the strings in Figure 21.7 to confirm that the hypercharge assignments in (21.60) and in (21.61) are correct.

We found more leptons than we wanted, but this is unavoidable in consistent models that only have D6-branes. There is a simple rule that must be obeyed.

Rule: the set of left-handed states that end on any collection of D-branes must contain equal numbers of incoming and outgoing strings. (21.62)

We will not derive this rule, but we can check that it holds for Figure 21.7. Consider the left branes, for example. There are nine outgoing quark doublets (three colors and three families). But there are also six incoming doublets from N_4 and three incoming doublets from N_5. This shows that we cannot obtain only three leptonic doublets.

Quick calculation 21.14 Verify that the numbers of ingoing and outgoing left-handed states are equal on each of the two right branes and on each of the two leptonic branes.

We can also state a general consistency condition that must be satisfied by the model. The condition has a topological character. Consider a point charge q at the origin of the (x, y) plane. In this two-dimensional world, the electric field lines extend radially from the charge all the way to infinity. On the other hand, if the plane is turned into a two-sphere of finite size, we have a problem with our charge: the field lines have nowhere to go. If the charge is at the north pole of the sphere, then the field lines will tend to go to the south pole. The solution is clear. Consistency requires a charge $(-q)$ at the south pole. The general conclusion is that the total charge in a compact space without boundary must be zero. This is easily derived.

Quick calculation 21.15 Consider the Maxwell equation $\nabla \cdot \vec{E} = \rho$, and integrate both sides of the equation over a compact space that has no boundary. Use the divergence theorem to show that the total charge must vanish.

In our present configuration we have D6-branes, which are objects that carry Ramond–Ramond (R–R) charge. Moreover, they wrap a boundaryless compact space T^6. So there

is a requirement of zero total charge. The formulation of the condition goes as follows. Introduce three pairs of formal variables: (x_1, y_1), (x_2, y_2), and (x_3, y_3). For any D6-brane a specified by

$$\ell_1^{(a)} = (m_1^a, n_1^a), \quad \ell_2^{(a)} = (m_2^a, n_2^a), \quad \ell_3^{(a)} = (m_3^a, n_3^a), \tag{21.63}$$

we construct the following polynomial of the six formal variables:

$$\Pi_a(x_i, y_i) \equiv (m_1^a\, x_1 + n_1^a\, y_1)\,(m_2^a\, x_2 + n_2^a\, y_2)\,(m_3^a\, x_3 + n_3^a\, y_3). \tag{21.64}$$

The polynomial Π_a encodes the R–R charges of the D6-brane of type a. If we expand the product, the polynomial consists of eight independent monomials. The consistency condition simply states that the result of adding the polynomials associated with each of the D6-branes in the configuration must be zero. If N_a denotes the number of D6-branes of type a, the consistency condition takes the form

$$\sum_a N_a\, \Pi_a(x_i, y_i) = 0. \tag{21.65}$$

This gives a total of eight consistency conditions on the numbers N_a and on the windings of the various D6-branes. For example, the vanishing of the coefficient of $x_1x_2x_3$ gives

$$\sum_a N_a\, m_1^a\, m_2^a\, m_3^a = 0. \tag{21.66}$$

This equation is easily tested using (21.51), (21.54), and (21.4). We find that it requires the vanishing of $N_1 - N_2 - N_3 - N_4 - 2N_5 + 3N_6$. Indeed, this quantity vanishes: $3 - 2 - 1 - 1 - 2 + 3 = 0$. In fact, the rule (21.4) is a consequence of (21.65), as you will show in Problem 21.2.

In order to build a string model with the particle content of the Standard Model, one needs to introduce orientifold planes, which effectively introduce image D6-branes. Remarkably, it is possible to build a configuration of branes and orientifolds so that (1) the gauge group is the Standard Model gauge group (21.25) and all additional $U(1)$ factors disappear, and (2) the set of chiral fermions is precisely that of the Standard Model (21.50). Intersecting D6-brane models are the first string theory models that have given precisely the particle content of the Standard Model. A qualification is necessary: the closed string sector of the theory may contain additional unobserved particles and interactions.

We will not discuss these models in full detail, but in order to satisfy your curiosity we will show the brane configuration and make a few remarks (a number of consistency checks have been relegated to Problem 21.3). The set of intersecting branes is shown in Figure 21.8. Comparing with the model in Figure 21.7, we see that there are still $N_1 = 3$ baryonic branes and $N_2 = 2$ left branes, but there is only one right brane $N_3 = 1$ and only one leptonic brane $N_4 = 1$. The orientifolds introduce image D-branes, which are denoted by affixing an asterisk to the brane labels and are shown as dashed lines. The configuration is chosen such that no brane intersects with its own image.

The main purpose of this construction is to avoid the extra leptonic doublets that we obtained above. In this model the left-handed quark doublets arise in a new way. The doublet in the first generation appears at the intersection $I_{12} = 1$ of the baryonic branes and

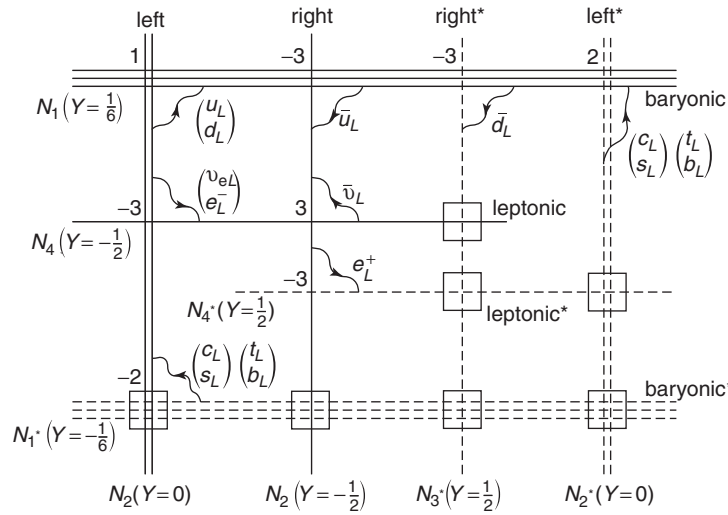

Fig. 21.8 **The brane configuration that leads to a Standard Model gauge group and matter content. There are $N_1 = 3$ baryonic branes, $N_2 = 2$ left branes, $N_3 = 1$ right branes and $N_4 = 1$ leptonic branes. The image D-branes created by the orientifolds are shown in dashed lines. Intersections framed by a square are mirrors of previously accounted intersections and do not give new particles.**

the left branes. The other two doublets appear at the intersection $I_{12^*} = 2$ of the baryonic branes and the image of the left branes (top right of the figure). Equivalently, the rules of orientifolds allow us to view these two doublets as appearing at the intersection $I_{1^*2} = -2$ of the left branes with the images of the baryonic branes (bottom left of the figure). Rule (21.4) continues to hold for the left branes: three quark doublets and three lepton doublets are outgoing, and six quark doublets are incoming.

We have implicitly been working with supersymmetric type IIA superstring theory; we are using stable D6-branes in a ten-dimensional spacetime. The branes and their intersections break all the supersymmetry. This had to be so, since we obtained the Standard Model spectrum, which does not have supersymmetry (the bosons and fermions do not appear in pairs of equal mass). In models without supersymmetry it is typically assumed that the string scale is low, perhaps on the order of a few TeV. This requires some large extra dimensions, otherwise the four-dimensional Planck length cannot be reproduced. In models with a low string scale, the lifetime of the proton can turn out to be too short. No proton decay has yet been observed, and the proton lifetime certainly exceeds 4×10^{33} years. In intersecting brane models the proton cannot decay by conventional processes where open strings join or split (see Problem 25.8), so proton decay appears to be safely suppressed.

You may ask: can we declare victory and state that the Standard Model has emerged from string theory? Not quite. The model must still do a lot more to be realistic. In particular, electroweak symmetry breaking must work out correctly. Recall that symmetry breaking is the process by which the Standard Model gauge group is reduced to (21.27), and the

fermions acquire mass. Moreover, the brane configuration is not fixed uniquely by the requirement that the Standard Model spectrum emerges; there are discrete choices to be made in selecting the brane wrapping numbers. In addition, there are other choices that must be made. For example, the positions of some D-branes on T^6 can be adjusted without altering the massless spectrum. The sizes of the tori are not automatically fixed either. If no choices had been possible, we would have a single candidate for the Standard Model within this class of brane configurations.

In intersecting D-brane models, electroweak symmetry breaking happens through a process of brane recombination. For certain values of the intersection angles between two branes, tachyon states appear in the spectrum of open strings that stretch between them. These tachyons indicate an instability that can cause the intersecting branes to combine at the intersection and produce a single brane. With fewer branes present, the gauge group is reduced. Moreover, the total number of intersections will be reduced and, consequently, the number of massless fermions will be reduced, as well (Problem 21.4). It is not clear whether any of these models will be able to produce the expected spectrum *after* symmetry breaking. This is not easy: a large number of mass parameters and other couplings must work out correctly. If this were possible, it would be a truly exciting result. There is a lot more that one can say about intersecting brane models and the more recent work that focuses on models that incorporate supersymmetry. It is now time, however, to discuss briefly other ways in which the Standard Model may appear in string theory.

21.5 String theory models of particle physics

Intersecting brane models are particularly attractive, but they are by no means the only avenues that have been investigated in order to construct string models of particle physics. We had a detailed look at the intersecting brane models because they are simple enough and can be understood very concretely. Past and present-day attempts to obtain the Standard Model follow several lines of attack, which can be organized by the starting points of the constructions. These possible starting points are the five supersymmetric string theories and M-theory (see Section 14.1). Each of these theories can be used to investigate how the Standard Model could emerge. The intersecting brane models that we discussed, for example, use type IIA superstrings as the starting point. Since the various supersymmetric strings and M-theory are different limits of a single theory, the various approaches are no doubt related at some level. Nevertheless, each starting point gives somewhat different insights into the various phenomenological questions that arise in the process of constructing a model.

The early attempts to do string phenomenology were based on the heterotic $E_8 \times E_8$ superstring theory. In this theory, six of the nine spatial dimensions are curled up into a small six-dimensional compact space with rather special properties: a *Calabi–Yau* space. Calabi–Yau spaces have both discrete and continuous parameters, which determine the details of the four-dimensional theory that arises upon compactification. For all Calabi–Yau spaces the minimal amount of supersymmetry survives the compactification; the resulting four-dimensional theory is said to be $N = 1$ supersymmetric. If supersymmetry is present, it must be minimal for chiral fermions to exist. The compactification also allows one to

break the original gauge symmetry $E_8 \times E_8$ down to $E_6 \times E_8$. One then views the E_8 factor as part of a hidden sector that has only indirect effects on our visible sector. The gauge group E_6 appears in the visible sector as the gauge group of a grand unified theory. The number of generations of chiral fermions depends on the topology of the Calabi–Yau space, and models with three generations can be obtained. The group E_6 contains $SU(3) \times SU(2) \times U(1)$ as a subgroup, so the Standard Model gauge group can arise upon further symmetry breaking. Calabi–Yau compactification of heterotic strings gave the first string models with semi-realistic particle physics. Since Calabi–Yau spaces are quite complicated, it turned out to be difficult to make progress with the questions of symmetry breaking and supersymmetry breaking.

The technical complications that must be faced with Calabi–Yau spaces prompted physicists to search for alternative six-dimensional spaces to compactify the heterotic string. If one simply uses a six-torus T^6, more than minimal supersymmetry survives and it becomes impossible to construct realistic models. Orbifolds provide a nice middle ground between Calabi–Yau spaces and tori. Orbifolds built from tori, for example, are much easier to analyze than general Calabi–Yau spaces, but they can still give $N = 1$ supersymmetric theories in four dimensions. It is possible to obtain semi-realistic models from orbifold compactifications of the heterotic string. Orbifolds also play an important role in the compactifications of other string theories.

More recently, physicists have been vigorously investigating the phenomenological possibilities of type II and type I superstring theories. One important class of models uses intersecting D-branes and orientifolds, as we considered in the previous section. Variations exist: one can use branes of different dimensionalities, or one can attempt to preserve some supersymmetry. A popular approach breaks the problem of constructing a fully consistent model into problems that can be analyzed separately.

One considers, for example, a type II string theory where the six extra dimensions form an infinite volume space that is everywhere flat except for a singularity at the origin. As you can imagine, flat extra dimensions with infinite volume cannot possibly give a realistic model: the extra dimensions would be visible. This complication, however, is ignored in the first part of the analysis. An example of such space is provided by the orbifold $\mathbb{C}^3/\mathbb{Z}_3$. Here $\mathbb{C}^3$ is six-dimensional flat space, viewed as the product of three copies of the complex plane, and $\mathbb{Z}_3$ describes the character of the identification that creates a singularity at the origin. Additionally, D3-branes are placed at the singularity. All the spatial directions on the branes lie along the four-dimensional space. It turns out that D-branes at orbifold singularities can give rise to gauge fields and chiral fermions. Models with a Standard Model-like spectrum do exist.

The second part of the analysis is a study of how to modify the non-compact six-dimensional space away from the singularity in order to turn it into a compact space of finite volume. If this is done in a way that leaves the region near the singularity unchanged, then the earlier results are preserved. Closing off the non-compact space may typically require the addition of other D-branes and/or orientifolds. A type II theory with D-branes and orientifolds can be analyzed alternatively as a type I superstring with D-branes. A type I superstring is a theory of unoriented open and closed strings. It can be viewed as the result of introducing a space-filling orientifold into type IIB string theory, which truncates the spectrum down to the subspace of states that are invariant under string orientation

reversal. While this truncation does not give a consistent string theory, the inconsistencies can be cured by adding open string degrees of freedom. The result is type I superstrings. The perspective of type I theory can be quite valuable in model building.

Finally, there are models based on M-theory. These are sometimes closely related to models built with type IIA superstrings, because M-theory compactified on a circle is type IIA superstring theory with some finite value for the string coupling. In fact, intersecting D6-brane models are closely related to M-theory models. From the viewpoint of M-theory, which is a theory in eleven dimensions, realistic physics requires compactification on a seven-dimensional manifold. In order to obtain four-dimensional theories with $N = 1$ supersymmetry, the seven-dimensional manifold must have G_2 holonomy, a geometrical property that is effectively a constraint on the curvature of the space. Chiral fermions and a reasonable gauge group can appear if the seven-dimensional space has singularities. The M-theory approach may give valuable insight into the effects of finite string coupling. Another popular approach that gives semi-realistic models uses a Calabi–Yau space to reduce the spacetime dimensionality to five. The fifth dimension is then turned into a finite segment, a space that can be viewed as an orbifold (Problem 2.5). This is called the heterotic M-theory approach, because the compactification of M-theory on a segment has been shown to give the heterotic $E_8 \times E_8$ superstring.

In summary, while a fully realistic model of particle physics has not yet been built in string theory, consistent progress towards this goal has been made. As we have seen, there are string models on D-branes whose open strings yield the particle content of the Standard Model. The significance of this development will depend on the ultimate success or failure of the models and what we learn from them. The intersecting brane models are not fully realistic. Symmetry breaking, for example, remains to be worked out. It would be a major accomplishment to achieve correct electroweak symmetry breaking in any string model.

If symmetry breaking works out in detail in some consistent string model, we would have shown that the Standard Model in its full glory can occur as a solution of string theory. Such a string model (or models?) might make interesting predictions that can be tested by new experiments. We would not be done, however. String theory is a theory of all the interactions, and it includes gravity. Quite a few other features have to work out. One notorious problem is that of stabilization of closed string moduli. Another is that of the cosmological constant. Finally, we need to make sure that the Standard Model is embedded into a consistent cosmology, which presumably includes inflation. We turn now our attention to these issues, where there has been some interesting progress.

21.6 Moduli stabilization and the landscape

In our study of T-duality of closed strings we encountered the simplest example of a moduli space of compactifications. One spatial coordinate was curled up into a circle whose radius R was arbitrary. The moduli space is simply the space of possible choices of the modulus R. The fact that the parameter R is a modulus means that the potential $V(R)$ for R vanishes

and, consequently, no particular value of R is selected. We can certainly choose a value of R and work with it but there is a complication. The theory in the reduced spacetime contains a massless scalar associated with fluctuations in the value of R. If we let x denote the coordinates of the reduced spacetime, the radius $R(x)$ of the circle at x is a massless field because the potential $V(R(x))$, constrained to vanish for constant $R(x)$, cannot contain a mass term $\frac{1}{2}m^2R^2(x)$. Unless its interactions are unnaturally small, a massless scalar field is inconsistent with observation. It follows that realistic compactifications of string theory must have no moduli.

By choosing a physical setup in which nontrivial potentials arise for would-be moduli, we can force them to acquire specific values. Fluctuations around those values now cost energy and thus represent allowed massive scalars. In a typical setup, would-be moduli include parameters of the compact space, positions of D-branes, and the value of the closed string dilaton. The goal of moduli stabilization is to provide potentials that fix the values of all moduli so that we have an admissible string background or, as is usually called, a string vacuum. While this seemed extremely difficult to achieve for a long time, developments in the period 2002–2005 led to the construction of *flux* compactifications in which all moduli are stabilized. Fluxes, to be described below, allow the stabilization of moduli but lead to an extremely large number of vacua. This *landscape* of vacua is so vast that it becomes plausible that in some of them the cosmological constant takes the value that we observe in nature – an extraordinarily small value in Planckian units. Flux compactification has also allowed some exploration of cosmology, in particular, the identification of mechanisms that would implement an inflationary period.

To develop insight into the issue of moduli stabilization let us discuss a six-dimensional theory of Einstein's gravity and electromagnetism. In this theory we curl up two spatial dimensions into a compact space so that we get a reduced four-dimensional spacetime M_4. We can consider metrics of the form

$$-ds^2 = g_{\mu\nu}(x)\,dx^\mu dx^\nu + R^2(x)\,\bar{g}_{ab}(y)dy^a dy^b. \tag{21.67}$$

The first term on the right-hand side is the metric on M_4, with $\mu, \nu = 0, \ldots, 3$. The second term represents the metric of the compact two-dimensional space with coordinates y^1 and y^2 – the indices a, b can take values one and two. In this metric we separated out a scale factor $R(x)$ that depends only on M_4 and a fixed metric $\bar{g}_{ab}(y)$ that depends only on the coordinates of the two-dimensional space. The volume of the two dimensional space is $R^2(x)V_2$ where V_2 is the volume computed with the metric $\bar{g}_{ab}(y)$. We thus see that the ansatz (21.67) represents correctly a compact space whose volume can fluctuate in the reduced spacetime. Deriving the potential $V(R)$ associated with R is a straightforward but technical calculation in Einstein's general relativity. We will only motivate the result. The potential is proportional to a topological invariant of the two-dimensional manifold: the Euler number $\chi = 2 - 2g$, where g is a non-negative integer called the genus. The two-sphere S^2, for example, has genus zero and Euler number two. A two-dimensional torus has genus one and Euler number zero. The genus of a compact two-dimensional surface is g if the surface is topologically a sphere with g holes. The potential $V(R)$ is not just a constant but it also has some R dependence. This dependence arises in the process of

disentangling the dynamics of gravity from that of R. As we have seen before, the dimensionally reduced gravitational constant G is equal to the higher-dimensional gravitational constant $G^{(6)}$ divided by the volume of the compact space. Since this volume can fluctuate, we seem to reach the paradoxical situation that G is not constant. Restoring a constant G requires a redefinition of the four-dimensional metric that introduces a factor of $1/R^4$ into the potential we are after. All in all, one obtains:

$$V(R) = -a_g \frac{\chi}{R^4}, \quad \chi = 2 - 2g, \tag{21.68}$$

where $a_g > 0$ is a constant. For a two-sphere $g = 0$ and the potential is negative definite $V \sim -1/R^4$. This means that left to its own devices a sphere will shrink to zero size. For a two-torus $g = 1$, the potential vanishes and R is a modulus. This is consistent with our experience with string theory T-duality where it is possible to curl up coordinates into circles of arbitrary radii; the space formed by two circles is a torus. For a space of $g > 1$, we have $V(R) \sim 1/R^4$ and left to its own devices the space will expand without limit. A compact space of genus zero has net positive curvature, a compact space of genus one has zero net curvature, and a compact space of genus greater than one has net negative curvature (this will be discussed further in §22.4). We thus see that positive net curvature induces collapse, negative net curvature induces blow up and zero net curvature gives us a modulus. None of these three situations is satisfactory, so we must consider extra elements that can help us stabilize the radius R.

If gravity is coupled to electromagnetism we can use magnetic flux to stabilize the radius. The magnetic flux on a torus is in fact quantized, as we learned in Section 19.3 and Problem 19.3. Assuming flux $\Phi = 2\pi n$, with n integer, we can easily estimate the contribution to the potential due to the magnetic field. If the characteristic area of the two dimensional space with the flux is R^2 the magnetic field magnitude goes like $B \sim n/R^2$. Since magnetic energy density is proportional to B^2 we get a total potential energy that goes like $R^2 B^2 \sim n^2/R^2$. With fixed flux the energy is reduced by making the space larger thus giving the space a tendency to expand. Due to the gravitational subtlety discussed above (21.68), the factor of $1/R^4$ also affects this contribution and the potential energy from the flux contributes to $V(R)$ a term that goes like n^2/R^6. We thus have

$$V(R) = -a_g \frac{\chi}{R^4} + a_f \frac{n^2}{R^6}, \tag{21.69}$$

where $a_f > 0$ is a constant. It is now clear that magnetic flux can stabilize a sphere: for $\chi = 2$ the potential above has a stable minimum for some $R > 0$. For different values n of the magnetic flux, we get different values of the critical radius. We have thus obtained an infinite family of stable solutions.

Quick calculation 21.16 Sketch the potential (21.69) for $\chi = 2$, determine the critical value of R, and confirm that it yields the minimum of the potential.

If we had wanted to fix R when it is a modulus ($\chi = 0$) and the first term in the potential (21.69) vanishes, in addition to magnetic flux we would need a negative contribution to $V(R)$. In string theory such negative contributions can arise from orientifolds. In a more

complete superstring compactification the extra dimensions form a six-dimensional space and we would have contributions to the potential from orientifolds, D-branes, and fluxes. If we take, for example, type IIB superstrings, the NS-NS field $B_{\mu\nu}$ and the RR field $A_{\mu\nu}$ have three-index field strengths that can provide fluxes. Just like magnetic fluxes are integrals of two-index field strengths. F_{ij} over two-dimensional manifolds, NS-NS and RR fluxes are integrals of three-index field strengths over three-dimensional manifolds. A typical six-dimensional Calabi–Yau space used for compactification can easily have hundreds of independent, non-contractible, three-dimensional submanifolds. Each one can support a flux characterized by an integer. As a result the fluxes on the Calabi–Yau can generate a potential that depends on hundreds of integers. A compactification of six dimensions into a Calabi–Yau with fluxes is an example of a *flux* compactification.

Faced with potentials arising from many fluxes, statistical considerations can help understand the space of possible vacua. As a very simplified but still illustrative example, we consider a potential for a single field ϕ characterized by two integers m and n:

$$V_{m,n}(\phi) = n\phi + \frac{1}{2}\, m\phi^2, \quad m, n \in \mathbb{Z}. \tag{21.70}$$

As m and n vary we get a set of potentials. In order to analyze this set concretely we limit the possible integers by the conditions

$$n^2 + m^2 \leq L, \quad m > 0. \tag{21.71}$$

The first constraint, with L a large integer, makes the set of potentials finite. The second constraint, $m > 0$, guarantees that the potentials have a stable minimum by making them bounded from below. Indeed, the critical point ϕ_* of (21.70) occurs at

$$0 = \frac{dV_{m,n}}{d\phi} = n + m\phi_* \quad \rightarrow \quad \phi_* = -\frac{n}{m}. \tag{21.72}$$

The value of ϕ_* represents the vacuum state in the potential $V_{m,n}$. We can also check that in this vacuum state the vacuum energy $\Lambda_{m,n}$ is

$$\Lambda_{m,n} = V_{m,n}(\phi_*) = -\frac{1}{2}\, \frac{n^2}{m}. \tag{21.73}$$

If we consider the set of all integers m, n allowed by (21.71), we can ask what is the approximate distribution of vacua ϕ_* in the set. We write

$$d\mathcal{N} = \rho(\phi_*)\, d\phi_*, \tag{21.74}$$

where $d\mathcal{N}$ is the number of critical points in the interval $(\phi_*, \phi_* + d\phi_*)$ and $\rho(\phi_*)$ is the distribution function we are after. We can represent the set of potentials as a set of points with integer coordinates on the (x, y) plane; each point (m, n) represents the potential $V_{m,n}(\phi)$ and the associated vacuum state ϕ_* (Figure 21.9). Because of (21.71) the allowed points lie to the right of the vertical axis and within a circle of radius $\sqrt{L}$ centered at the origin. For large L the total number of points, or total number $\mathcal{N}$ of vacua, is approximately equal to the area of the allowed region: $\mathcal{N} \simeq \frac{1}{2}\pi L$. We also note that $|\phi_*| = \frac{|n|}{m} < \sqrt{L}$ since $m \geq 1$, and $n < \sqrt{L}$.

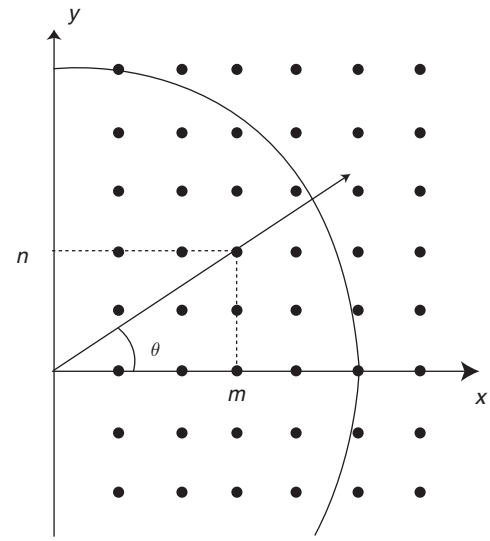

Fig. 21.9 A set of potentials $V_{m,n}(\phi)$ defined by two integers m, n, with $m > 0$.

To determine the distribution of vacua, we note that $\phi_* = -\frac{n}{m}$ implies that points on a fixed radial line give the same vacuum value. In fact

$$\phi_* = -\tan\theta \quad \text{or} \quad \theta = -\tan^{-1}\phi_*, \tag{21.75}$$

where θ is the angle that the radial line makes with respect to the positive x-axis. The number of vacua $d\mathcal{N}$ in the small sector $(\theta, \theta + d\theta)$ is equal to the area

$$d\mathcal{N} = \frac{L}{2}|d\theta| = \frac{L}{2}\,\frac{1}{1+\phi_*^2}\,|d\phi_*| \quad \rightarrow \quad \rho(\phi_*) = \frac{L}{2}\,\frac{1}{1+\phi_*^2}, \tag{21.76}$$

where we used (21.75) to relate $d\theta$ to $d\phi_*$ and (21.74) to read the value of $\rho(\phi_*)$. The distribution has the expected property $\rho(-\phi_*) = \rho(\phi_*)$, which arises because with $\phi_* = -n/m$, for each critical point with a given $n = n_0$ there is another with $n = -n_0$. On the given set, the distribution $\rho(\phi_*)$ applies for $|\phi_*| < \sqrt{L}$. The shape of ρ indicates that the density of vacua peaks at $\phi_* = 0$. In the limit of very large L we confirm that the number $\mathcal{N}$ of vacua is

$$\mathcal{N} = \frac{L}{2}\int_{-\sqrt{L}}^{\sqrt{L}} \frac{d\phi_*}{1+\phi_*^2} \simeq \frac{L}{2}\int_{-\infty}^{\infty} \frac{d\phi_*}{1+\phi_*^2} = \frac{1}{2}\pi L. \tag{21.77}$$

Another interesting distribution is that of vacuum energies (21.73). Letting $d\mathcal{N}$ denote the number of vacua with energy in the interval $(\Lambda, \Lambda + d\Lambda)$ we write

$$d\mathcal{N} = \rho(\Lambda)\,|d\Lambda|, \tag{21.78}$$

where $\rho(\Lambda)$ is the associated distribution function. A calculation (Problem 21.6) gives

$$\rho(\Lambda) = \frac{4}{3}\,\frac{1}{\sqrt{2|\Lambda|}}\left[\sqrt{\Lambda^2 + L} - |\Lambda|\right]^{3/2}. \tag{21.79}$$

Since all vacuum energies in this example are negative (see (21.73)), the above distribution must be supplemented with $\rho(\Lambda) = 0$ for $\Lambda > 0$. We note that $\rho(\Lambda)$ grows as $\Lambda \to 0$. For further discussion see Problem 21.6.

Statistical considerations are relevant when we discuss how string theory compactifications could reproduce the observed value of the so-called dark energy of the universe. This dark energy is the present value of the vacuum energy and, most likely, can be identified with a cosmological constant. While vacuum energy presently dominates the energy density of the universe, it is extremely small in Planckian units. To see this explicitly we first note that assuming that the Hubble expansion parameter takes the value $H_0 = 73.5\,\text{km/s/Mpc}$, the present critical energy density ρ_c of the universe is

$$\rho_c = \frac{3H_0^2}{8\pi G} = 1.015 \times 10^{-29}\ \text{g/cm}^3. \tag{21.80}$$

This density is equivalent to that of about six protons per cubic meter.

Quick calculation 21.17 Consider the natural Planckian mass density $\rho_P = m_P/\ell_P^3$. Show that $\rho_P = 5.20 \times 10^{93}\ \text{g/cm}^3$.

Current experimental evidence indicates that the energy density of the universe is very close to the critical density and, apparently, vacuum energy contributes as much as 76% of it. We therefore find that the vacuum energy density is

$$\rho_{\text{vac}} \simeq 0.76\,\rho_c = 0.771 \times 10^{-29}\frac{\text{g}}{\text{cm}^3} = 1.48 \times 10^{-123}\,\rho_P. \tag{21.81}$$

The present vacuum energy is about 123 orders of magnitude smaller than the natural value in Planckian units! Following conventional string theory notation, we use the symbol Λ to denote vacuum energy density. We do not follow the notation in general relativity where Λ is the "cosmological constant," a quantity obtained by multiplying the vacuum energy by $8\pi G/c^4$. We thus write (21.81) as

$$\Lambda_{\text{obs}} \simeq 1.48 \times 10^{-123}\,\Lambda_P. \tag{21.82}$$

Since string units are usually assumed to be only a couple of orders of magnitude different from Planckian units, the emergence of an extraordinarily small vacuum energy is a puzzle. Even in scenarios with large extra dimensions, Λ_{obs} would still be many orders of magnitude smaller than the natural string value. In string theory compactifications one expects vacuum energies in the range $(-\Lambda_P,\ \Lambda_P)$. If only a small number of string vacua existed, it would seem very unlikely that any of them would have an energy equal to $\sim 10^{-123}\Lambda_P$. As we will argue now, in a flux compactification with many fluxes the number of vacua is so large that it becomes possible to find a sizable number of vacua with energy in the experimentally allowed window. Consider then the potential for a set of moduli represented by a vector $\vec{\phi}$ in a theory with J fluxes characterized by integers $n_1, n_2, \ldots, n_J$:

$$V(\vec{\phi}\,) = V_0(\vec{\phi}\,) + \sum_{i=1}^{J} m_i(\vec{\phi}\,)\ n_i^2. \tag{21.83}$$

Here $V_0(\vec{\phi})$ is the part of the potential created by effects other than fluxes. It is the analog of the first term on the right-hand side of (21.69). The second term in (21.83) is the contribution from the fluxes and involves calculable functions $m_i(\vec{\phi})$. This term is the analog of the second term on the right-hand side of (21.69). When the number of moduli and the number J of fluxes are both large it becomes difficult to find the critical points of the potential. Following a strategy that was vindicated by more detailed analysis, we simply freeze the moduli $\vec{\phi}$ at some typical value $\vec{\phi}_0$ and skip the minimization process. Both V_0 and the functions m_i in (21.83) become just constants. Writing $\Lambda = V$ and $\Lambda_0 = V_0$ we then find,

$$\Lambda = \Lambda_0 + \sum_{i=1}^{J} q_i^2 n_i^2. \tag{21.84}$$

We have written $m_i = q_i^2$ to make it manifest that fluxes contribute positively to the energy density. Moreover, as in typical compactifications, Λ_0 will be taken to be a negative, Planckian size energy density:

$$\Lambda_0 \sim -\Lambda_{\rm P}. \tag{21.85}$$

We want to understand if there is a reasonable number of vacua for which the total vacuum energy is in the observational window, say between zero and $\Delta\Lambda$ where

$$\Delta\Lambda \simeq \Lambda_{\rm obs} \simeq 10^{-123}\Lambda_{\rm P}. \tag{21.86}$$

For the vacua in this interval we have

$$0 \leq \Lambda_0 + \sum_{i=1}^{J} q_i^2 n_i^2 \leq \Delta\Lambda. \tag{21.87}$$

Since Λ_0 is negative we write this as

$$|\Lambda_0| \leq \sum_{i=1}^{J} q_i^2 n_i^2 \leq |\Lambda_0| + \Delta\Lambda. \tag{21.88}$$

We want to know how many vacua satisfy this inequality. The number would be large if the q_i^2 were small compared to $\Delta\Lambda$. The problem is that the q_i^2 are *many* orders of magnitude larger than $\Delta\Lambda$. They are, in fact, Planckian, so at best a few orders of magnitude smaller than $|\Lambda_0|$.

If a vacuum characterized by nonvanishing integers $(n_1, n_2, \ldots, n_J)$ satisfies (21.88), we have at least 2^J degenerate vacua $(\pm n_1, \pm n_2, \ldots, \pm n_J)$ that do. If the q_i satisfy special relations further degeneracies can occur: if $q_i = q_j$, for example, the exchange of n_i with n_j yields additional degenerate vacua. We will call $d(J) \simeq 2^J$, the degeneracy of vacua.

To analyze the constraint (21.88) efficiently we introduce Cartesian coordinates

$$x_i = q_i n_i, \tag{21.89}$$

so that vacua become lattice points in x-space. Since lattice points are separated by distances q_i along x_i, the cells that contain a single lattice point have a volume

$$\text{vol}_{\text{cell}} = \prod_{i=1}^{J} q_i . \tag{21.90}$$

In x-space the constraint (21.88) becomes

$$|\Lambda_0| \le \sum_{i=1}^{J} x_i^2 \le |\Lambda_0| + \Delta\Lambda . \tag{21.91}$$

This is a spherical shell in x-space. To find the thickness dR of the shell we write $|\Lambda_0| = R_0^2$ and $|\Lambda_0| + \Delta\Lambda = (R_0 + dR)^2$ from which we find

$$dR = \frac{\Delta\Lambda}{2\sqrt{|\Lambda_0|}} . \tag{21.92}$$

The x-volume of the shell is equal to the volume of the sphere S^{J-1} of radius $R_0 = \sqrt{|\Lambda_0|}$ times the thickness dR:

$$\text{vol}_{\text{shell}} = \text{vol}(S^{J-1}) \sqrt{|\Lambda_0|}^{J-1} \frac{\Delta\Lambda}{2\sqrt{|\Lambda_0|}} , \quad \text{vol}(S^{J-1}) = \frac{\pi^{J/2}}{\Gamma(J/2)} , \tag{21.93}$$

where we recalled (3.52). Since vacua are $d(J)$-fold degenerate, we have vacua in the shell if the number of lattice points in the shell, divided by $d(J)$, is greater than or equal to one:

$$\frac{1}{d(J)} \frac{\text{vol}_{\text{shell}}}{\text{vol}_{\text{cell}}} = \frac{\text{vol}(S^{J-1})}{2d(J)} \sqrt{|\Lambda_0|}^{J-1} \frac{\Delta\Lambda}{\sqrt{|\Lambda_0|}} \frac{1}{\prod_{i=1}^{J} q_i} \ge 1 . \tag{21.94}$$

A rearrangement gives

$$\frac{\Delta\Lambda}{2|\Lambda_0|} \ge \frac{d(J)}{\text{vol}(S^{J-1})} \prod_{i=1}^{J} \frac{q_i}{\sqrt{|\Lambda_0|}} . \tag{21.95}$$

We want the ratio to the left of the inequality to be about 10^{-124}. For $J = 1$, ignoring the order one prefactor gives the unnaturally small $q \sim 10^{-124}\sqrt{|\Lambda_0|}$. For large J we need to pay some attention to the prefactor, which gets large. Indeed, $d(J) \simeq 2^J$ grows with J and $\text{vol}(S^{J-1})$ becomes smaller with J.

Quick calculation 21.18 Verify that the volume $\text{vol}(S^k)$ reaches a maximum for $k = 6$ and decreases monotonically for larger k.

In order to perform the large J analysis of (21.95) we use the asymptotic expansion

$$\text{vol}(S^{J-1}) \simeq \left(\frac{2\pi e}{J}\right)^{J/2} , \quad J \to \infty . \tag{21.96}$$

Moreover, we set the left-hand side of (21.95) equal to 10^{-124}, write $\prod_{i=1}^{J} q_i = q^J$, introducing an effective q, and set $d(J) = 2^J$. This gives

$$10^{-124} \ge \left(\frac{2\pi e}{J}\right)^{-J/2} \left(\frac{2q}{\sqrt{|\Lambda_0|}}\right)^{J} . \tag{21.97}$$

and leads to

$$\frac{2q}{\sqrt{|\Lambda_0|}} \simeq 10^{-\frac{124}{J}} \sqrt{\frac{2\pi e}{J}} \equiv f(J). \tag{21.98}$$

For $J = 124$, for example, one gets $f(J) = 0.037$, a number that shows that qs just a little smaller than the natural Planckian value suffice to get string vacua in the experimentally allowed window.

Quick calculation 21.19 Use $\Gamma(x) \sim x^{x-1}\sqrt{2\pi x}\, e^{-x}(1 + \mathcal{O}(1/x))$ to prove (21.96).

Quick calculation 21.20 The function $f(J)$ attains a maximum for J near six-hundred. Determine the critical J and the corresponding value of $f(J)$.

We have produced, via flux compactification, a sizable number of vacua with energies in the observable window. Of course, there are many more vacua with unacceptably large energies. The argument does not tell us why the vacuum energy is so small, it just shows that it *can* be that small. If string theory incorporates inflation an unusual kind of explanation has been offered. Inflation, as presently understood, is eternal into the future. With eternal inflation, bubble universes nucleate forever and, most likely, all vacua in the landscape are eventually realized as physical bubble universes. This includes universes like ours, in which the vacuum energy is small. We find ourselves in such universe, it is said, because had the vacuum energy been substantially larger, galaxies and stars could not have formed and life could not exist. However outrageous this argument seems, it has had some success. Before its experimental discovery in 1999, most theoreticians believed that the vacuum energy was exactly zero. Every attempt to use theories to prove it had to be zero, however, had failed. Weinberg showed that structure formation in the universe gave an upper bound on the vacuum energy. He also argued that if no principle guaranteed a vanishing vacuum energy it would probably turn out to have a value smaller than the bound, but not dramatically smaller. This turned out to be correct.

Since the experimental evidence for a period of inflation has become much stronger in the last few years, one can ask if inflation occurs in string theory. Most investigations have been done using flux compactifications since moduli must be stabilized. Typically, one imagines a Calabi–Yau space with an elongation that has an anti-D-brane at its tip. A D-brane slowly moving along the elongation towards the anti-D-brane may produce inflation. It is too early to tell if these efforts will lead to complete models.

As one contemplates the current efforts to show that string theory describes the real world one cannot fail to experience mixed feelings. On the one hand, string models and flux compactifications are evidence that string theory *is* promising, after all, central features of the real world can emerge from the theory. On the other hand, the constructions seem contrived, at least in the sense that they are engineered to give the physics that we observe, rather than obtained naturally as the simplest solutions of string theory. The Standard Model of particle physics and the Standard Cosmological Model are intricate constructs, and present-day attempts to describe them within string theory do not result in simplification. Despite this

complexity, you need not react as Alfonso X (the Wise, 1221–1284) appears to have done when faced with Ptolemy's astronomy:[1]

> Had I been present at the creation, I would have given some useful hints for the better ordering of the universe.

It is conceivable that the landscape idea is correct and our universe is just one out of a gigantic number of possibilities, one that is not especially natural. This would explain why the string constructions do not seem natural. At this stage, however, it may be more constructive to continue to explore the facts and the models, looking for naturalness to emerge as our understanding improves. This is based on the hope that, as stated by Maimonides (1135–1204) in *The Guide for the Perplexed* (1190),

> In the realm of Nature there is nothing purposeless, trivial, or unnecessary.

Problems

Problem 21.1 Oblique tori and orientifold action.

Consider the (x, y) plane with an orientifold line along the x axis. The orientifolding symmetry operation reverses string orientation (this will not play a role here) and acts on the plane by sending every point (x, y) into its image $(x, -y)$. The action is well defined because it takes points on the plane to points on the plane. The orientifold line is a line which is fixed under the orientifold action.

A two-torus is obtained by implementing two identifications on the (x, y) plane. If the identifications are those in (2.101) and (2.102) the result is a square torus. To get a rectangular torus we simply use

$$(x,\ y) \sim (x + a,\ y), \quad (x,\ y) \sim (x,\ y + b), \quad a, b > 0. \tag{1}$$

The fundamental domain of the identifications together with its boundary can be taken to be the region whose points (x, y) satisfy $0 \leq x \leq a$ and $0 \leq y \leq b$. This is a rectangle with sides a and b. Draw a picture illustrating the situation.

(a) Convince yourself that the orientifold action $(x, y) \to (x, -y)$ is well defined on the torus (1). On the torus, the line $0 \leq x \leq a$, $y = 0$ is a fixed line. There is another fixed line on the torus. Find it. Show that the orientifold action on the torus can be viewed as a reflection about either one of the two fixed lines. So, there are really two orientifold lines on the torus!

Consider now the class of oblique two-tori that are obtained by implementing the identifications

$$(x,\ y) \sim (x + a_1,\ y + a_2), \quad (x,\ y) \sim (x,\ y + b), \quad a_1, a_2, b > 0. \tag{2}$$

[1] I am grateful to J. Goldstone for providing this quote from Bartlett's *Familiar Quotations* (1919). Alfonso X was Spanish king of Castile and Leon. Moses Maimonides was born in Cordova, Spain. The intersecting D6-brane models we examined in Section 21.4 were developed by the "Spanish group."

The fundamental domain of the identifications together with its boundary can be taken to be the parallelogram $\mathcal{P}$ with a vertex at the origin and sides defined by the vectors $\vec{a} = (a_1, a_2)$ and $\vec{b} = (0, b)$. Let $\vec{x} = (x, y)$, so that the identifications can be written as $\vec{x} \sim \vec{x} + \vec{a}$ and $\vec{x} \sim \vec{x} + \vec{b}$. Draw a picture showing the parallelogram $\mathcal{P}$.

(b) Consider now the torus (2) when

$$a_2 = b/2. \tag{3}$$

Use an appropriate figure to convince yourself that the orientifold action $(x, y) \to (x, -y)$ is well defined on the torus. Find the two fixed lines on $\mathcal{P}$. Note that these lines represent a *single* closed orientifold line on the torus.

(c) Let $\vec{a}^* \equiv (a_1, -a_2)$. Show that (3) implies that $\vec{a}^* = \vec{a} - \vec{b}$. Note the significance of this relation on your sketch. Let points be denoted by vectors $t_1\vec{a} + t_2\vec{b}$, with constants t_1 and t_2. Show that the orientifold action takes

$$t_1\,\vec{a} + t_2\vec{b} \;\longrightarrow\; t_1\,\vec{a} - (t_1 + t_2)\,\vec{b}. \tag{4}$$

(d) A closed line ℓ on the torus can be denoted by a pair of relatively prime integers $\ell = (m, n)$ with the understanding that the line is obtained from the straight segment between the origin and the point $m\vec{a} + n\vec{b}$. Examine the argument that established (21.5) for the intersection number on a square torus. Convince yourself that this formula also holds for lines on oblique tori. It follows from (4) that after orientifold action

$$\ell = (m, n) \;\longrightarrow\; \ell^* = (\,m,\; -m - n\,). \tag{5}$$

It is convenient to encode lines with another pair of numbers, one of which can be half-integral:

$$\ell = (m, n) \;\text{ is represented by }\; \ell = [\,m, n + \tfrac{m}{2}\,]. \tag{6}$$

(e) Show that the orientifold action takes

$$\ell = [\,r,\; s\,] \;\longrightarrow\; \ell^* = [\,r, -s\,]. \tag{7}$$

(f) Show that the intersection number for two lines $\ell_1 = [r_1, s_1]$ and $\ell_2 = [r_2, s_2]$ is

$$\#(\ell_1, \ell_2) = r_1 s_2 - r_2 s_1. \tag{8}$$

Show that

$$\#(\ell_1^*, \ell_2^*) = -\#(\ell_1, \ell_2). \tag{9}$$

(g) Orient the orientifold closed line on the torus from left to right. Show that it can be described as a vector of the form $t_1\vec{a} + t_2\vec{b}$ that is encoded as [2, 0].

Problem 21.2 Intersection numbers and formal variables.

In Section 21.4 we introduced formal variables x_i, y_i, $i = 1, 2, 3$ to describe a polynomial Π that encodes the charge of a D6-brane wrapped on a T^6. Consider now the following rules for products of the formal variables:

$$\begin{aligned} &x_i\, y_j = y_j\, x_i, \quad x_i\, x_j = x_j\, x_i, \quad \text{and} \quad y_i\, y_j = y_j\, y_i, \quad \text{for} \quad i \neq j \\ &x_i\, x_i = y_i\, y_i = 0, \quad \text{and} \quad x_i\, y_i = -y_i\, x_i = 1, \quad i \text{ not summed.} \end{aligned} \tag{1}$$

Moreover, assume that the product obeys the distributive law. Note that the product is not commutative (the order of the factors matters) and that the above rules cannot be used to simplify the polynomial Π associated with a brane.

(a) Show that the product of the polynomials associated with two wrapped D6-branes gives their intersection number: $\Pi_a \Pi_b = I_{ab}$.
(b) Use (21.65) to derive the consistency condition $\sum_a N_a I_{ab} = 0$, for any fixed b. Explain in detail why this condition gives the rule formulated in (21.62).

Problem 21.3 An intersecting brane model with the particle content of the Standard Model.

To solve this problem you must first solve Problem 21.1, which deals with orientifolds on oblique tori. The T^6 in the present brane model is composed of two rectangular tori (the first and the second tori) and one oblique torus (the third torus). On the oblique torus, we denote lines as $[p, q]$.

On each rectangular torus, the wrapping of the two fixed orientifold lines is represented by $(2, 0)$. On the oblique torus, the single orientifold line is represented by $[2, 0]$ (Problem 21.1, part (g)). As a result, the T^6 carries three of the spatial directions along the O6-planes. The other three directions along the O6-planes coincide with the spatial directions of the effective four-dimensional spacetime.

The four sets of D6-branes of the model are described by the lines

$$\begin{aligned} N_1 = 3: &\quad \ell_1^{(1)} = (1, 0), \quad \ell_2^{(1)} = (5, 1), \quad \ell_3^{(1)} = [\,1, 1/2], \\ N_2 = 2: &\quad \ell_1^{(2)} = (0, -1), \quad \ell_2^{(2)} = (1, 0), \quad \ell_3^{(2)} = [\,1, 3/2], \\ N_3 = 1: &\quad \ell_1^{(3)} = (4, 3), \quad \ell_2^{(3)} = (1, 0), \quad \ell_3^{(3)} = [\,0, 1], \\ N_4 = 1: &\quad \ell_1^{(4)} = (1, 0), \quad \ell_2^{(4)} = (1, 1), \quad \ell_3^{(4)} = [1, -3/2]. \end{aligned} \tag{1}$$

The mirror image i^* of a D6-brane i is obtained by changing the sign of the second entry on each of the three lines that define the brane.

(a) Give the (m, n) values corresponding to the standard description of each of the four lines on the third torus.
(b) Verify that $I_{ii^*} = 0$ for $i = 1$, 2, 3, and 4. This ensures that no D6-brane intersects its own mirror image.
(c) Calculate all other intersection numbers I_{ij} and I_{ij^*}, with $i < j$. Verify that

$$\begin{aligned} &I_{12} = 1, \;\; I_{12^*} = 2\,, \;\; I_{13} = -3, \;\; I_{13^*} = -3, \;\; I_{14} = I_{14^*} = 0, \\ &I_{23} = I_{23^*} = 0, \;\; I_{24} = -3\,, \;\; I_{24^*} = 0, \;\; I_{34} = 3, \;\; I_{34^*} = -3. \end{aligned} \tag{2}$$

Confirm that Figure 21.8 correctly represents the information encoded by the above intersection numbers.

(d) Since the orientifold wrappings are described by $(2, 0)$, $(2, 0)$, $[2, 0]$, the polynomial (21.64) takes the form $\Pi_{O6} = 8x_1x_2x_3$. Since the Ramond–Ramond charge of an orientifold O6-plane is opposite to that of a D6-brane and four times larger, the condition of vanishing total R–R charge on T^6 is

$$\sum_{i=1}^{4} N_i \, (\Pi_i + \Pi_{i^*}) - 4 \, \Pi_{O6} = 0. \tag{3}$$

Verify that the wrapping numbers in (1) satisfy this constraint. Note: the last factor in the polynomial Π is built using the numbers in the $[p, q]$ representation.

(e) In this model, one of the four $U(1)$ factors remains massless. Its charge is proportional to $Q_1 - 3(Q_3 + Q_4)$. Here $Q_i = +1$ for a string ending on an i-brane or beginning on an i^*-brane. Consider the states in Figure 21.8. Verify that, with suitable normalization, the charge is precisely the hypercharge.

Problem 21.4 Symmetry breaking by recombination of intersecting branes.

In order to simplify our work, consider a pair of intersecting D4-branes. Each brane has one spatial direction wrapped as a closed line on a square torus T^2 with coordinates x^4 and x^5. The other three spatial directions on the world-volume of the branes stretch along the spatial directions of the four-dimensional spacetime. The details of the compactification of $x^6, \ldots, x^9$ are not relevant.

Since we have a single T^2, the wrapping of each D4-brane can be specified by a single line $\ell = (m, n)$, where m and n are relatively prime. Given two D4-branes with lines $\ell^{(1)} = (m_1, n_1)$ and $\ell^{(2)} = (m_2, n_2)$, the intersection number I_{12} which determines the number of left-handed fermions at the intersection set is simply given by

$$I_{12} = \#(\ell^{(1)}, \ell^{(2)}) = m_1n_2 - m_2n_1.$$

The process of brane recombination can be described using the corresponding lines on the torus. Take one of the branes, say the first one, and cut out an infinitesimal piece near the *endpoint* of its line. The result is an open line on the torus beginning at $(0, 0)$ and ending near $(0, 0)$. Take the second brane, and cut an infinitesimal piece near its *beginning*. The result is an open line beginning near $(0, 0)$ and ending at $(0, 0)$. Now glue the end of the first line with the beginning of the second. The result is a single closed line representing a recombined D-brane.

(a) Use the representation of the defining lines on the plane (where the torus arises as the unit square) to argue that the recombination of two D-branes $\ell^{(1)} = (m_1, n_1)$ and $\ell^{(2)} = (m_2, n_2)$ gives a D-brane represented by a line that begins at $(0, 0)$, ends at $(m_1 + m_2, n_1 + n_2)$, and has one corner (where?). The line can be deformed continuously into a straight line joining the endpoints. This line, which represents the recombined D-brane Σ, is $\ell_\Sigma = (m_1 + m_2, n_1 + n_2)$. Explain why the final result of recombination does not depend on the order in which the D-branes are glued together.

(b) Carry out explicitly the recombination procedure on the torus for the branes $(1, 0)$ and $(0, 1)$. Show in a sequence of figures how the recombined branes can be continuously deformed into the brane $(1, 1)$.

(c) When two branes recombine, the total number of branes decreases by one unit and, as a result, the number of Maxwell fields decreases by one unit. Consider a D-brane i which intersects each of the two branes, brane 1 and brane 2, that recombine into Σ. Prove that

$$|I_{i\Sigma}| \leq |I_{i1}| + |I_{i2}|.$$

This result shows that the number of (massless) chiral fermions obtained after brane recombination is less than or equal to the total number of chiral fermions that were present before recombination. We thus have the expected signs of symmetry breaking: a number of Maxwell fields and a number of chiral fermions acquire mass.

Problem 21.5 Distribution of vacuum energies in toy model.

Derive the distribution $\rho(\Lambda)$ given in (21.79). One possible approach is to consider the continuous approximation in which $(n, m) = (x, y)$ and thus $\Lambda = -\frac{1}{2}\frac{y^2}{x}$ on the allowed space. Then calculate the area in between the curves associated with Λ and $\Lambda + d\Lambda$.

What is the upper bound for $|\Lambda|$ in the set of allowed vacua? Test the consistency of (21.79) by using it to calculate the total number of vacua. Take $L = 100$ and estimate the fraction of vacua that have $|\Lambda| < 1$.

Problem 21.6 Distribution of vacuum energies with large number of fluxes.

Consider equation (21.84) with $\prod_{i=1}^{J} q_i = q^J$ and J is large so that (21.96) applies.

(a) Determine the number $\mathcal{N}(\Lambda)$ of vacua with vacuum energy less than or equal to Λ. The answer depends on Λ and the parameters $|\Lambda_0|$, q, and J.

(b) Use differentiation to write an equation of the form

$$d\mathcal{N} = \rho(\Lambda)\,\frac{d\Lambda}{|\Lambda_0|},$$

and determine the distribution $\rho(\Lambda)$. The distribution depends parametrically on $|\Lambda_0|$, q, and J.

(c) Assume $J = 100$ and $q = 0.01\sqrt{|\Lambda_0|}$. Determine the value of $\rho(\Lambda = 0)$. Find the value of Λ_* (in terms of $|\Lambda_0|$) for which $\rho(\Lambda_*)$ differs from $\rho(0)$ by 1%. Calculate the ratio $\Lambda_*/\Lambda_{\text{obs}}$. Your result must demonstrate that the distribution $\rho(\Lambda)$ is accurately constant over ranges of conceivable experimental relevance. How many string vacua are there with Λ values within 1% of Λ_{obs}?

22 String thermodynamics and black holes

The thermodynamics of strings is governed largely by the exponential growth of the number of quantum states accessible to a string, as a function of its energy. We estimate such growth rates by counting the number of partitions of large integers. The behavior of the entropy indicates that at high energies the temperature approaches a finite constant, the Hagedorn temperature. The finite temperature single-string partition function for open bosonic strings is calculated. We explain how the counting of string states can be used to give a statistical mechanics derivation of the entropy of black holes. The calculations give results in qualitative agreement with the entropy of Schwarzschild black holes and in quantitative agreement with the entropy of certain charged black holes.

22.1 A review of statistical mechanics

Our study of string thermodynamics will make use of both the microcanonical and canonical ensembles. Recall that the *microcanonical ensemble* consists of a collection of copies of a particular system A, one for each state accessible to A at a particular fixed energy E. In the *canonical ensemble* we consider the system A in thermal contact with a reservoir at a temperature T. This ensemble contains copies of the system A together with the reservoir, one copy for each allowed state of the combined system. In the canonical ensemble the energy of system A varies among members of the ensemble.

Let us begin with the microcanonical ensemble. The system A is imagined to be in isolation and to have a fixed energy. We let $\Omega(E)$ denote the number of possible states of the system A when it has energy E. The entropy S of the system is defined in terms of the number of states as

$$S(E) = k \ln \Omega(E), \tag{22.1}$$

where k is Boltzmann's constant. The temperature T of the system is defined in terms of the derivative of the entropy with respect to the energy:

$$\frac{1}{T} = \frac{\partial S}{\partial E}. \tag{22.2}$$

It is sometimes easier to work with the canonical ensemble. Imagine a system A which has a fixed volume and which is in thermal contact with a reservoir at temperature T. This system could be a box full of strings, or it could be a box containing a single string. It is not necessary to specify what the reservoir is. Suppose we know the quantum states $\{\alpha\}$ of

the system and their associated energies $\{E_\alpha\}$. Then the partition function Z for system A is defined as

$$Z \equiv \sum_\alpha e^{-\beta E_\alpha}, \quad \beta = \frac{1}{kT}. \tag{22.3}$$

The partition function is useful because it can be used to calculate interesting quantities. For instance, if system A is known to have temperature T, then, using Z, we can calculate the probability that A is in a particular quantum state. By definition, the partition function depends both on the temperature T and on the external parameters of the system. These are the parameters that determine the energy levels of the system. The systems we will consider have only one external parameter: the volume V occupied by the system. Thus we will think of Z as $Z(T, V)$, or

$$Z = Z(\beta, V). \tag{22.4}$$

The probability P_α that the system, in contact with the reservoir at temperature T, is in the state α is

$$P_\alpha = \frac{e^{-\beta E_\alpha}}{Z}. \tag{22.5}$$

Clearly $\sum_\alpha P_\alpha = 1$, as required by the interpretation of P_α as a probability. We can calculate the average energy E of the system A in the ensemble by differentiation of the partition function:

$$E = \sum_\alpha P_\alpha E_\alpha = -\frac{\partial \ln Z}{\partial \beta}. \tag{22.6}$$

The pressure p of the system can also be calculated from the partition function (Problem 22.1). It is given by

$$p = \frac{1}{\beta}\frac{\partial \ln Z}{\partial V}. \tag{22.7}$$

Another useful quantity is the Helmholtz free energy F. Its basic properties can be obtained in a few steps starting from the first law of thermodynamics. This law states that the energy change dE in a system whose only external parameter is the volume V can be written as

$$dE = TdS - pdV. \tag{22.8}$$

Here T is the temperature of the system and p is the pressure. Moreover, TdS is the heat transferred into the system, and $(-pdV)$ is the mechanical work done *on* the system. Equation (22.8) implies that E should be viewed as a function $E(S, V)$ of S and V, and,

$$T = \left(\frac{\partial E}{\partial S}\right)_V, \quad p = -\left(\frac{\partial E}{\partial V}\right)_S. \tag{22.9}$$

We can also write the change in energy in (22.8) as

$$dE = d(TS) - S\,dT - p\,dV, \tag{22.10}$$

which means that

$$d(E - TS) = -S\,dT - p\,dV. \tag{22.11}$$

The free energy F is defined as

$$F \equiv E - TS, \tag{22.12}$$

and therefore we have

$$dF = -S\,dT - p\,dV. \tag{22.13}$$

We see that for processes at constant temperature the free energy represents the amount of energy that can go into mechanical work. For a chemical reaction that releases energy, for example, the entropy of the system typically decreases. Not all of the energy released can then be used for work; only the free energy can. Since the total entropy cannot decrease, the rest of the energy goes into heat, which increases the entropy of the world. It follows from (22.13) that F should be viewed as a function $F(T, V)$ of T and V, and,

$$S = -\Big(\frac{\partial F}{\partial T}\Big)_V, \quad p = -\Big(\frac{\partial F}{\partial V}\Big)_T. \tag{22.14}$$

The free energy can be calculated from the partition function (Problem 22.1). It is given by

$$F = -kT \ln Z. \tag{22.15}$$

Our aim is to use the basic thermodynamic relations reviewed above to compute interesting properties of the string. One central computation is that of the partition function for a string. This problem is a bit complex, so we first consider simpler problems that will help us build the necessary tools.

The first result we need is a formula for the number of partitions of large integers. We will obtain this mathematical result using a physical method: the analysis of the high-temperature behavior of a quantum *nonrelativistic* string, call it a "quantum violin string." With this result, we calculate the entropy/energy relation for an *idealized* quantum *relativistic* string, a string for which we ignore the momentum labels of the quantum states. The Hagedorn temperature already emerges in this context. After a discussion of the partition function for the relativistic point particle, we assemble all of our results to compute the partition function of the relativistic string.

In the latter part of this chapter we discuss a significant success of string theory: a statistical mechanics derivation of the entropy of black holes. This entropy, first arrived at via thermodynamical considerations, arises from the degeneracy of string states that have the macroscopic properties of black holes. The agreement between the string calculations and the thermodynamical expectation is only qualitative for the case of Schwarzschild black holes, but it is quantitative for certain types of extremal black holes.

22.2 Partitions and the quantum violin string

Consider a quantum mechanical nonrelativistic string with fixed endpoints: a quantum violin string. This string, studied classically in Chapter 4, has an infinite set of vibrating frequencies, all multiples of a basic frequency ω_0. Its idealization as a quantum string is a collection of simple harmonic oscillators with frequencies $\omega_0, 2\omega_0, 3\omega_0$, and so on. Each simple harmonic oscillator (SHO) has its own creation and annihilation operators, as well as its own Hamiltonian:

$$\begin{aligned}
&\text{SHO}_{\omega_0}: \quad (a_1, a_1^\dagger)\,, \quad H_{\omega_0} = \hbar\omega_0\, a_1^\dagger a_1\,,\\
&\text{SHO}_{2\omega_0}: \quad (a_2, a_2^\dagger)\,, \quad H_{2\omega_0} = 2\,\hbar\omega_0\, a_2^\dagger a_2\,,\\
&\text{SHO}_{3\omega_0}: \quad (a_3, a_3^\dagger)\,, \quad H_{3\omega_0} = 3\,\hbar\omega_0\, a_3^\dagger a_3\,,\\
&\quad\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\quad \vdots
\end{aligned} \tag{22.16}$$

Here we have discarded zero point energies, and all oscillators satisfy the conventional commutation relations

$$[a_m, a_n^\dagger] = \delta_{mn}. \tag{22.17}$$

Since the quantum string is the union of all these oscillators, the Hamiltonian $\hat{H}$ is

$$\hat{H} = \sum_{\ell=1}^{\infty} H_{\ell\omega_0} = \hbar\omega_0 \sum_{\ell=1}^{\infty} \ell\, a_\ell^\dagger a_\ell. \tag{22.18}$$

We recognize here the number operator $\hat{N}$:

$$\hat{H} = \hbar\omega_0\, \hat{N}\,, \quad \hat{N} = \sum_{\ell=1}^{\infty} \ell\, a_\ell^\dagger a_\ell. \tag{22.19}$$

The vacuum state of the string is a state $|\Omega\rangle$ such that

$$a_\ell|\Omega\rangle = 0\,, \quad \text{for all } \ell. \tag{22.20}$$

A quantum state $|\Psi\rangle$ of this string is obtained by letting creation operators act on the vacuum:

$$|\Psi\rangle = (a_1^\dagger)^{n_1}(a_2^\dagger)^{n_2}\cdots(a_l^\dagger)^{n_l}\cdots|\Omega\rangle\,. \tag{22.21}$$

The state is therefore specified by the set $\{n_1, n_2, n_3, \ldots\}$ of occupation numbers. The number operator acting on the state $|\Psi\rangle$ gives us

$$\hat{N}|\Psi\rangle = N|\Psi\rangle, \tag{22.22}$$

where

$$N = n_1 + 2n_2 + 3n_3 + \cdots = \sum_{\ell=1}^{\infty} \ell n_\ell. \tag{22.23}$$

Table 22.1 Counting states of fixed total number eigenvalue N; $p(N)$ denotes the number of partitions of the integer N

N	List of states	$p(N)$
1	$a_1^\dagger$	1
2	$a_2^\dagger\,,\ (a_1^\dagger)^2$	2
3	$a_3^\dagger\,,\ a_2^\dagger a_1^\dagger\,,\ (a_1^\dagger)^3$	3
4	$a_4^\dagger\,, a_3^\dagger a_1^\dagger\,, (a_2^\dagger)^2\,, a_2^\dagger (a_1^\dagger)^2\,, (a_1^\dagger)^4$	5

It then follows from (22.19) that the energy E of $|\Psi\rangle$ is given by

$$E = \hbar\omega_0\, N. \tag{22.24}$$

A counting question arises naturally here. For a fixed positive integer N, how many states are there with $\hat{N}$ eigenvalue equal to N? This number, denoted by $p(N)$, is so important that it has been given a name: the number of *partitions* of N. Before explaining the reason for this terminology, let us determine $p(N)$ for $N = 1, 2, 3,$ and 4. Shown in Table 22.1 are the states with those values of N. For brevity, we show only the oscillators, omitting the vacuum state $|\Omega\rangle$ that they act on. The fourth line, for example, shows that there are five states with $\hat{N}$ eigenvalue equal to four. Thus $p(4) = 5$.

It is appropriate to name the quantity $p(N)$ the number of partitions of N. A partition of N is a set of positive integers that add up to N. The order of the elements in the set is immaterial. Thus, for example, {3, 2} is a partition of 5, and so is {2, 1, 1, 1}. The partitions of 4 are

$$\{4\}\,,\quad \{3,1\}\,,\quad \{2,2\}\,,\quad \{2,1,1\}\,,\quad \{1,1,1,1\}. \tag{22.25}$$

The number of states with $\hat{N}$ eigenvalue equal to N coincides with the number of partitions of N. Indeed, given a partition of N we can build a state by attaching each element of the partition as a subscript to an oscillator $a^\dagger$ and letting the resulting collection of oscillators act on the vacuum. Note that this is exactly how the states in the last line of Table 22.1 are built from the partitions of 4 given in (22.25). Conversely, given a state with number N, the set of subscripts of all the oscillators in the state gives a partition of N.

We would like to find a formula for $p(N)$, but our analysis will not give us that much. We will derive an expression that describes $\ln p(N)$ accurately for large N. A more refined calculation gives the famous approximation for $p(N)$ found by Hardy and Ramanujan.

Our strategy will be as follows. We know that the entropy S is given as a function of the energy E by (22.1). For a given E, $N = E/(\hbar\omega_0)$, and $\Omega(E)$ is simply $p(N)$. Therefore

$$S(E) = k \ln\ p(N) = k \ln\ p\left(\frac{E}{\hbar\omega_0}\right). \tag{22.26}$$

If we find $S(E)$ then we will have found the function $p(N)$. To find $S(E)$ we will calculate the partition function Z for the quantum violin string. From Z, we will find the free energy F. We will be able to evaluate the free energy explicitly only in the case of high temperature. It will then be easy to find the high energy behavior of the entropy $S(E)$. This we will use to find a large-N approximation for $p(N)$.

Let us now begin with the calculation of the partition function. We have

$$Z = \sum_{\alpha} \exp\left(-\frac{E_\alpha}{kT}\right) = \sum_{n_1,n_2,n_3,\ldots} \exp\left[-\frac{\hbar\omega_0}{kT}(n_1 + 2n_2 + 3n_3 + \cdots)\right]. \tag{22.27}$$

In writing this equation we have recognized that the set of all states is labeled by the set of occupation numbers. To sum over all states is to sum over all occupation numbers, each of which ranges from zero to infinity. Since the exponential of a sum can be written as a product of exponentials, the sums over different occupation numbers can be made independently,

$$Z = \sum_{n_1} \exp\Big[-\frac{\hbar\omega_0}{kT}\, n_1\Big] \cdot \sum_{n_2} \exp\Big[-\frac{\hbar\omega_0}{kT}\, 2n_2\Big] \cdots. \tag{22.28}$$

Therefore we have

$$Z = \prod_{\ell=1}^{\infty} \sum_{n_\ell=0}^{\infty} \exp\left(-\frac{\hbar\omega_0 \ell n_\ell}{kT}\right). \tag{22.29}$$

The sum over each n_ℓ is a geometric series, so we find

$$Z = \prod_{\ell=1}^{\infty} \left[1 - \exp\left(-\frac{\hbar\omega_0 \ell}{kT}\right)\right]^{-1}. \tag{22.30}$$

Finally, the free energy F is found using (22.15):

$$F = -kT \ln Z = kT \sum_{\ell=1}^{\infty} \ln\Big[1 - \exp\Big(-\frac{\hbar\omega_0 \ell}{kT}\Big)\Big]. \tag{22.31}$$

We cannot go any further unless we make some approximations. If the temperature T is high enough so that

$$\frac{\hbar\omega_0}{kT} \ll 1, \tag{22.32}$$

then each term in the sum (22.31) differs very little from the previous one. This allows us to approximate the sum by an integral:

$$F \simeq kT \int_1^{\infty} d\ell\, \ln\Big[1 - \exp\Big(-\frac{\hbar\omega_0 \ell}{kT}\Big)\Big]. \tag{22.33}$$

The choice $\ell = 1$ for the lower limit of integration, as opposed to any other finite, small number, plays no role. Indeed, changing the variable of integration to

$$x = \frac{\hbar\omega_0}{kT}\, \ell, \tag{22.34}$$

we find that the lower limit of integration becomes $x = 0$ in the high temperature limit. As a result, we obtain

$$F \simeq \frac{(kT)^2}{\hbar\omega_0}\int_0^\infty dx\,\ln(1-e^{-x}). \tag{22.35}$$

Using the expansion

$$\ln(1-y) = -\Big(y + \frac{1}{2}y^2 + \frac{1}{3}y^3 + \frac{1}{4}y^4 + \cdots\Big), \tag{22.36}$$

which is valid for any $0 \le y < 1$, we have

$$\begin{aligned} F &\simeq -\frac{(kT)^2}{\hbar\omega_0}\int_0^\infty dx\left(e^{-x} + \frac{1}{2}e^{-2x} + \frac{1}{3}e^{-3x} + \frac{1}{4}e^{-4x}\cdots\right) \\ &\simeq -\frac{(kT)^2}{\hbar\omega_0}\Big[1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots\Big]. \end{aligned} \tag{22.37}$$

The sum in brackets is a familiar one. It is, in fact, the zeta function (12.109) with an argument of two:

$$\zeta(2) = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots = \frac{\pi^2}{6}. \tag{22.38}$$

Thus we finally obtain the high temperature approximation for the free energy:

$$F \simeq -\frac{(kT)^2}{\hbar\omega_0}\frac{\pi^2}{6} = -\frac{1}{\hbar\omega_0}\frac{\pi^2}{6}\frac{1}{\beta^2}. \tag{22.39}$$

Evidently, for this string the free energy has no volume dependence.

We can now calculate the entropy as a function of temperature. Using (22.14) we find

$$S = -\frac{\partial F}{\partial T} = k\,\frac{\pi^2}{3}\left(\frac{kT}{\hbar\omega_0}\right). \tag{22.40}$$

Since we are interested in the entropy as a function of energy, we also compute the energy. Making use of (22.6) we have

$$E = -\frac{\partial \ln Z}{\partial\beta} = \frac{\partial}{\partial\beta}(\beta F) = -\frac{\pi^2}{6}\frac{1}{\hbar\omega_0}\frac{\partial}{\partial\beta}\left(\frac{1}{\beta}\right), \tag{22.41}$$

which gives

$$E = \frac{\pi^2}{6}\frac{1}{\hbar\omega_0}\frac{1}{\beta^2} = \frac{\pi^2}{6}\left(\frac{kT}{\hbar\omega_0}\right)^2\hbar\omega_0. \tag{22.42}$$

Quick calculation 22.1 Verify that the energy E can also be calculated from $F = E - TS$.

Combining (22.40) and (22.42) yields

$$S(E) = k\pi\sqrt{\frac{2}{3}\frac{E}{\hbar\omega_0}} = k\,2\pi\sqrt{\frac{N}{6}}. \tag{22.43}$$

Table 22.2 Comparing the exact values of $p(N)$ with the estimate $p(N)_{\text{est}}$ provided by the Hardy–Ramanujan formula

N	$p(N)$	$p(N)_{\text{est}}$	$p(N)/p_{\text{est}}(N)$
5	7	8.94	0.7829
10	42	48.10	0.8731
100	190 569 292	199 281 893.25	0.9563
1000	2.406×10^{31}	2.440×10^{31}	0.9860
10 000	3.617×10^{106}	3.633×10^{106}	0.9956

Comparing with equation (22.26), we finally read

$$\ln p(N) \simeq 2\pi\sqrt{\frac{N}{6}}. \tag{22.44}$$

This was our goal, an estimate of $\ln p(N)$ for large N. Indeed, we must require large N since

$$N = \frac{E}{\hbar\omega_0} = \frac{\pi^2}{6}\left(\frac{kT}{\hbar\omega_0}\right)^2 \gg 1, \tag{22.45}$$

because of our high temperature assumption (22.32).

The result (22.44) is only the leading term of the celebrated Hardy–Ramanujan asymptotic expansion of $p(N)$:

$$p(N) \simeq \frac{1}{4N\sqrt{3}} \exp\left(2\pi\sqrt{\frac{N}{6}}\right). \tag{22.46}$$

This is not an exact formula either, but it is an accurate estimate of $p(N)$, as opposed to our accurate estimate of the logarithm of $p(N)$. We will not give here a derivation of the Hardy–Ramanujan result. It is fun, however, to test the accuracy of the Hardy–Ramanujan formula. In Table 22.2 we compare the values of $p(N)$, as calculated exactly, with the estimate $p_{\text{est}}(N)$ provided by (22.46). The estimate gives an error of about one-half of one percent for $N = 10\,000$.

We now need a minor generalization of (22.46). Assume the string can vibrate in b transverse directions. Then for each frequency $\ell\omega_0$ we have b harmonic oscillators that represent the possible polarizations of the motion. The associated occupation numbers need a superscript to label the b polarizations:

$$\begin{matrix} n_1^{(1)} & n_1^{(2)} & \dots & n_1^{(b)} \\ n_2^{(1)} & n_2^{(2)} & \dots & n_2^{(b)} \\ \dots & \dots & \dots & \dots \\ n_\ell^{(1)} & n_\ell^{(2)} & \dots & n_\ell^{(b)} \\ \dots & \dots & \dots & \dots \end{matrix} \tag{22.47}$$

In order to sum over all possible states in the new partition function Z_b, we must sum over all possible values of the occupation numbers $n_k^{(q)}$, where $k = 1, 2, \dots, \infty$, and $q = 1, 2, \dots, b$. This gives

$$Z_b = \sum_{n_k^{(1)},\dots,n_k^{(b)}} \exp\Big[-\frac{\hbar\omega_0}{kT}\sum_{\ell=0}^{\infty}\sum_{q=1}^{b}\ell n_\ell^{(q)}\Big]. \tag{22.48}$$

The sums over the various $n^{(q)}$ factorize, so

$$Z_b = \sum_{n_k^{(1)}} \exp\Big[-\frac{\hbar\omega_0}{kT}\sum_{\ell=0}^{\infty}\ell n_\ell^{(1)}\Big]\cdots\sum_{n_k^{(b)}} \exp\Big[-\frac{\hbar\omega_0}{kT}\sum_{\ell=0}^{\infty}\ell n_\ell^{(b)}\Big]. \tag{22.49}$$

Each factor here is equal to the partition function Z calculated previously, so

$$Z_b = (Z)^b. \tag{22.50}$$

The new free energy F_b is also easy to calculate:

$$F_b = -kT\ln Z_b = -kTb\ln Z = bF. \tag{22.51}$$

The entropy, obtained by differentiation of the free energy, also acquires a multiplicative factor of b:

$$S_b = bS. \tag{22.52}$$

For the energy E_b, the same multiplicative factor exists on account of (22.6):

$$E_b = bE. \tag{22.53}$$

The four equations above are equalities of functions of temperature. For example, (22.52) is $S_b(T) = b\,S(T)$, where $S(T)$ is the entropy determined previously in (22.40). Since E is the previously determined energy (as a function of temperature), S and E are related by (22.43). Note that E_b is equal to $\hbar\omega_0 N$, where N is now the total occupation number

$$E_b = bE\ = \hbar\omega_0 N\,, \quad N = \sum_{\ell,q}\ell n_\ell^{(q)}. \tag{22.54}$$

Using (22.52), our earlier result for $S(E)$ in (22.43), and (22.54) we find

$$S_b = b\,(k\,2\pi)\sqrt{\frac{1}{6}\frac{E}{\hbar\omega_0}} = k\,2\pi\sqrt{\frac{b}{6}\frac{Eb}{\hbar\omega_0}} = k\,2\pi\sqrt{\frac{Nb}{6}}. \tag{22.55}$$

Let us call $p_b(N)$ the number of partitions of N into integers that carry any of b labels. This means, for example, that the partition $\{3, 2, 1\}$ of 6 now gives rise to many partitions

written as $\{3_{p_1}, 2_{p_2}, 1_{p_3}\}$, where the subscripts p_1, p_2, and p_3 can take any value from one to b. A partition with different subscripts is considered a different partition. We now see that with b degenerate oscillators, the number of states with total number N is $p_b(N)$. Therefore $S_b = k \ln p_b(N)$, and comparing with (22.55) we conclude that for large N

$$\ln\, p_b(N) \simeq 2\pi\sqrt{\frac{Nb}{6}}. \tag{22.56}$$

The more accurate version of this result can be shown to be

$$p_b(N) \simeq \frac{1}{\sqrt{2}}\left(\frac{b}{24}\right)^{(b+1)/4} N^{-(b+3)/4} \exp\left(2\pi\sqrt{\frac{Nb}{6}}\right). \tag{22.57}$$

You can see that for $b = 1$ this reduces to $p(N)$, as given in (22.46). For $b = 24$, the number of transverse light-cone directions in the bosonic string, the expression simplifies a little:

$$p_{24}(N) \simeq \frac{1}{\sqrt{2}}\, N^{-27/4} \exp\left(4\pi\sqrt{N}\right). \tag{22.58}$$

Quick calculation 22.2 Show that for large N

$$\frac{p_{24}(N+1)}{p_{24}(N)} \simeq \exp\left(\frac{2\pi}{\sqrt{N}}\right). \tag{22.59}$$

This means that the fractional change in the number of partitions when the argument is increased by one unit goes down to zero as $N \to \infty$.

Quick calculation 22.3 Use direct counting to confirm that $p_{24}(1) = 24$, $p_{24}(2) = 324$, $p_{24}(3) = 3200$, and $p_{24}(4) = 25\,650$.

It is also interesting to count other types of partitions. Consider, for example, partitions of integers into unequal parts. The possible partitions of 6 into unequal integers are

$$\{6\}, \quad \{5, 1\}, \quad \{4, 2\}, \quad \{3, 2, 1\}. \tag{22.60}$$

We denote by $q(N)$ the number of partitions of N into unequal parts, so $q(6)$, for example, is equal to four. We can use a fermionic version of the violin string to determine the large-N behavior of $q(N)$. The frequencies of the oscillators are not changed, but this time we demand that each occupation number can only be equal to zero or to one. Since no creation operator can be used more than once, the total number N of any state is effectively split into contributions all of whose parts are unequal. Creation operators that cannot be used more than once create fermionic excitations. We call such oscillators fermionic. With a little abuse of language, the numbers that enter a partition into unequal parts are called fermionic numbers. You will show in Problem 22.2 that, for large N,

$$\ln q(N) \sim 2\pi\sqrt{\frac{N}{12}}. \tag{22.61}$$

We extended the earlier counting of $p(N)$ to the case in which the elements of a partition can carry any of b labels. If the elements of an unequal partition can carry any of f labels, then the number $q_f(N)$ of partitions is obtained from (22.61) by replacing $N \to Nf$. In such partitions a fermionic number can appear more than once if it carries a different label each time. This counting corresponds to a system with f species of fermionic oscillators.

A final generalization is also useful. We consider partitions of N into both ordinary and fermionic numbers, with b labels for the ordinary numbers and f labels for the fermionic numbers. In this case (Problem 22.4) we find that the large-N leading behavior of the number $P(N; b, f)$ of such partitions is

$$\ln P(N; b, f) \simeq 2\pi \sqrt{\frac{N}{6}\Big(b + \frac{f}{2}\Big)}. \tag{22.62}$$

As an example, let us calculate $P(2; 1, 2)$, the number of partitions of 2 into ordinary and fermionic numbers, with the latter having two possible labels. The list of partitions is obtained by labeling the numbers in the ordinary partitions ($\{2\}$ and $\{1, 1\}$) in all possible ways. We find

$$\{2\},\ \{2_1\},\ \{2_2\},\ \{1\ ,\ 1\},\ \{1_1\ ,\ 1\},\ \{1_2\ ,\ 1\},\ \{1_1\ ,\ 1_2\}. \tag{22.63}$$

The labels on the fermionic numbers are shown as subscripts. We count $P(2; 1, 2) = 7$.

Equation (22.62) is useful for calculations in superstring theories, where the states are built with both bosonic and fermionic creation operators. An application to a supersymmetric black hole will be considered in Section 22.7.

22.3 Hagedorn temperature

Let us now return to the subject of relativistic strings. We will consider open strings that carry no spatial momentum. This will happen, for example, if the open string endpoints end on a D0-brane. With zero spatial momentum, the string has energy levels that are simply given by the rest masses of its quantum states. The mass-squared of a given state can be expressed in terms of the number operator $N^\perp$ (12.164):

$$M^2 = \frac{1}{\alpha'}(N^\perp - 1) \simeq \frac{N^\perp}{\alpha'}, \tag{22.64}$$

in the approximation of large $N^\perp$. It follows that the energy $E = M$ is related to the number operator by the simple equality

$$\sqrt{N^\perp} = \sqrt{\alpha'}\, E. \tag{22.65}$$

In the microcanonical ensemble, the number of states $\Omega(E)$ equals $p_{24}(N^\perp)$, because we have 24 transverse light-cone directions, and consequently 24 oscillator labels for each

mode number. As a result, $S(E) = k \ln p_{24}(N^\perp)$. For large energy, $N^\perp$ is also large, and using equation (22.56) we find

$$S(E) = k\, 2\pi \sqrt{\frac{N^\perp \cdot 24}{6}} = k\, 4\pi \sqrt{N^\perp}. \tag{22.66}$$

Making use of the number–energy relation in (22.65) we find

$$S = k\, 4\pi \sqrt{\alpha'}\, E. \tag{22.67}$$

This is the entropy–energy relation at high energy. An entropy proportional to the energy is unusual because it leads to a constant temperature:

$$\frac{1}{kT} = \frac{1}{k} \frac{\partial S}{\partial E} = 4\pi \sqrt{\alpha'}. \tag{22.68}$$

This temperature is called the Hagedorn temperature $T_{\rm H}$:

$$\frac{1}{\beta_{\rm H}} = kT_{\rm H} = \frac{1}{4\pi \sqrt{\alpha'}}. \tag{22.69}$$

Here $kT_{\rm H}$ is the thermal energy associated with the Hagedorn temperature. In the high energy approximation we are working with, we can arbitrarily increase the energy of the strings, and their temperature will remain fixed at the Hagedorn temperature. It is interesting to compare the energy $kT_{\rm H}$ to the rest mass of the particles found in the first massive level of the open string. This corresponds to $N^\perp = 2$ in (22.64) and gives $E = M = 1/\sqrt{\alpha'}$. The ratio of the Hagedorn thermal energy to this rest energy is

$$\frac{kT_{\rm H}}{\left(1/\sqrt{\alpha'}\right)} = \frac{1}{4\pi} \simeq \frac{1}{12.6}. \tag{22.70}$$

This shows that the Hagedorn thermal energy is small compared with the rest energy of almost any particle state of the string. This is an important result that will play a role in our later work in this chapter.

The entropy–energy relation in (22.67) also holds for closed strings with no spatial momentum. Recalling (13.48), we find

$$M^2 = \frac{2}{\alpha'}(N^\perp + \bar{N}^\perp - 2) \simeq \frac{4}{\alpha'} N^\perp, \tag{22.71}$$

since closed string states satisfy $N^\perp = \bar{N}^\perp$. It follows that the energy $E = M$ is related to the number operator as

$$2\sqrt{N^\perp} = \sqrt{\alpha'}\, E. \tag{22.72}$$

This time, the number of states $\Omega(E)$ is equal to the product of available states in the left-moving and in the right-moving sectors:

$$\Omega(E) = p_{24}\left(N^{\perp}\right) p_{24}\left(\bar{N}^{\perp}\right) = \left(p_{24}(N^{\perp})\right)^2 . \tag{22.73}$$

As a result, the entropy S is precisely twice that indicated in (22.66):

$$S(E) = k\, 4\pi\, (2\sqrt{N^{\perp}}) = k\, 4\pi\sqrt{\alpha'}E. \tag{22.74}$$

This is the same entropy–energy relation we had for open strings. We conclude that the temperature $T_{\rm H}$ is also the approximate temperature of highly energetic closed strings.

22.4 Relativistic particle partition function

As a warmup for our computation of the partition function for a string, we compute here the partition function for a particle. We will work with a relativistic particle of mass m that lives in a D-dimensional spacetime, or equivalently, in $d = D - 1$ space dimensions. Moreover, we assume that this particle is confined to a box of volume V:

$$V = L_1 L_2 \ldots L_d. \tag{22.75}$$

This box is in thermal contact with a reservoir at temperature T. As usual, the energy and the momentum of the particle are related by

$$E(\vec{p}) = \sqrt{\vec{p}^{\,2} + m^2}. \tag{22.76}$$

The quantum states of the particle in the box are labeled by the momenta $\vec{p}$, which are quantized. The partition function $Z(m^2)$ is given by

$$Z(m^2) = \sum_{\vec{p}} \exp(-\beta E(\vec{p}\,)). \tag{22.77}$$

The volume dependence of this partition function arises because the quantized values of the momenta depend on the dimensions of the box. The quantum wavefunctions with momentum $\vec{p} = \hbar\vec{k}$ have a spatial dependence $\exp(i\vec{k}\cdot\vec{x})$. The periodicity of these wavefunctions in the box requires that for each spatial direction i

$$k_i L_i = 2\pi n_i\,, \quad i = 1, 2, \ldots, d\,, \quad n_i \in \mathbb{Z}. \tag{22.78}$$

Equivalently, in terms of momenta,

$$n_i = p_i\, \frac{L_i}{(2\pi\hbar)}. \tag{22.79}$$

It follows that summing over the various momenta is the same as summing over the various n_i. For an arbitrary smooth function $f[E]$ of the energy we can thus write

$$\sum_{\vec{p}} f[E(\vec{p})] = \sum_{\vec{n}} f[E(\vec{p}(\vec{n}))] \simeq \int dn_1 dn_2 \ldots dn_d\, f[E(\vec{p}\,(\vec{n}))], \tag{22.80}$$

where the approximation by an integral is allowed because, for large boxes, the momenta change very little when a counter n_i shifts by one unit. Using (22.79) and (22.75) we obtain

$$\sum_{\vec{p}} f[E(\vec{p})] \simeq V \int \frac{d^d\vec{p}}{(2\pi\hbar)^d}\ f[E(\vec{p})]. \tag{22.81}$$

This is the general prescription for dealing with sums over momenta. Applied to our case of interest (22.77) it gives

$$Z(m^2) = V \int \frac{d^d\vec{p}}{(2\pi\hbar)^d} \exp\Big(-\beta\sqrt{\vec{p}^2+m^2}\Big). \tag{22.82}$$

This is the integral representation of the partition function for a relativistic point particle of rest mass m. The temperature and volume arguments of Z are implicit. Working with $\hbar = 1$, and changing variables of integration by letting $\vec{p} = m\vec{u}$, we find

$$Z(m^2) = Vm^d \int \frac{d^d\vec{u}}{(2\pi)^d} \exp\left(-\beta m\sqrt{1+\vec{u}^{\,2}}\right). \tag{22.83}$$

This integral is not elementary, but it can be written in terms of derivatives of modified Bessel functions with argument βm (Problem 22.6). Rather than doing so, we will examine the integral in the domain of interest. For our string theory applications, the thermal energy is much smaller than the rest energy of the particle. Indeed, as we saw earlier, for temperatures below the Hagedorn temperature, almost all string states satisfy this condition. Thus we consider the situation where

$$\beta m \gg 1\,, \quad \text{low temperature.} \tag{22.84}$$

We now claim that the leading approximation to the integral can be found by expanding the square root in (22.83) for $\vec{u}^{\,2}$ small. This is explained as follows. Using spherical coordinates and letting $\vec{u}^{\,2} = u^2$, we note that $d^d\vec{u} \sim u^{d-1}du$ (recall the familiar cases of $d = 2, 3$). As a plain one-dimensional integral, the integrand in (22.83) is thus of the form

$$\text{Integrand} \ \sim\ u^{d-1}\, e^{-\beta m\sqrt{1+u^2}}. \tag{22.85}$$

This integrand vanishes at $u = 0$ and $u = \infty$, and it peaks somewhere in between, giving the largest contribution to the integral. The maximum of the integrand can be found by setting the u derivative of (22.85) equal to zero. This gives the condition

$$\frac{d-1}{\beta m} = \frac{u^2}{\sqrt{1+u^2}}. \tag{22.86}$$

Since βm is large, the left-hand side is small, and u^2 must also be small. We can therefore neglect the u^2 in the square root, and we find that the integrand is largest for

$$u^2 \simeq \frac{d-1}{\beta m} \ll 1. \tag{22.87}$$

We are therefore allowed to expand the square root in (22.83) to write

$$Z(m^2) \simeq Vm^d e^{-\beta m} \int \frac{d^d\vec{u}}{(2\pi)^d} \exp\Big(-\frac{1}{2}\,\beta m\,\vec{u}^2\Big). \tag{22.88}$$

The integral is now Gaussian and is readily evaluated:

$$Z(m^2) \simeq V e^{-\beta m} \Big(\frac{m}{2\pi\beta}\Big)^{\frac{d}{2}}. \tag{22.89}$$

This is our final form for the partition function of a relativistic particle in the low temperature limit. One can verify that this partition function is dimensionless, as it should be. Except for the additional factor $e^{-\beta m}$, this partition function coincides with the exact partition function for a nonrelativistic particle. The exponential factor accounts for the contribution of the rest energy to the energy of the relativistic particle.

22.5 Single string partition function

We are now ready to evaluate the partition function for a single open string placed in a box of volume V. In order to calculate this, we must enumerate the quantum states of the string. The states are obtained by acting with the light-cone creation operators on the momentum eigenstates. A set of basis states is written as in (12.162):

$$|\lambda,\, p\rangle = \prod_{n=1}^{\infty} \prod_{I=2}^{25} (a_n^{I\dagger})^{\lambda_{n,I}}\, |p^+,\, \vec{p}_T\rangle, \tag{22.90}$$

where the notation $|\lambda,\, p\rangle$ emphasizes that the momentum components as well as the occupation numbers $\lambda_{n,I}$ are labels of the string states. The d components $(p^+, \vec{p}_T)$ listed in the momentum eigenstate specify the light-cone energy p^- via the on-shell condition:

$$M^2(\{\lambda_{n,I}\}) = -p^2 = 2p^+p^- - p^I p^I, \tag{22.91}$$

where

$$M^2(\{\lambda_{n,I}\}) = \frac{1}{\alpha'}(N^\perp - 1)\,, \quad N^\perp = \sum_{n,I} n\,\lambda_{n,I}. \tag{22.92}$$

Since both the spatial momentum and the energy are determined for the above states, we can label the string states with the set $\{\lambda_{n,I}\}$ of occupation numbers and the *spatial* momentum $\vec{p}$. We then write

$$E(\{\lambda_{n,I}\},\, \vec{p}\,) = \sqrt{M^2(\{\lambda_{n,I}\}) + \vec{p}^2}. \tag{22.93}$$

To find the partition function $Z_{\rm str}$ of a single string, we must sum over all states $|\lambda,\, p\rangle$, or equivalently, over all spatial momenta $\vec{p}$ and all values of the occupation numbers $\lambda_{n,I}$:

$$Z_{\rm str} = \sum_{\alpha} \exp(-\beta E_\alpha) = \sum_{\lambda_{n,I}} \sum_{\vec{p}} \exp\Big[-\beta\sqrt{M^2(\{\lambda_{n,I}\}) + \vec{p}^2}\Big]. \tag{22.94}$$

The momentum sum simply gives the partition function Z for a relativistic particle of mass-squared $M^2(\{\lambda_{n,I}\})$. We thus write

$$Z_{\rm str} = \sum_{\lambda_{n,I}} Z\,(M^2(\{\lambda_{n,I}\})). \tag{22.95}$$

Since the mass M^2 depends only on $N^\perp$, the sum over occupation numbers $\{\lambda_{n,I}\}$ can be traded for a sum over $N^\perp \equiv N$, as long as we keep in mind that there are $p_{24}(N)$ states with number eigenvalue N:

$$Z_{\rm str} = \sum_{N=0}^{\infty} p_{24}(N)\, Z(M^2(N)). \tag{22.96}$$

So far, no approximations have been made, and the above result is exact.

Let N_0 denote an integer for which $p_{24}(N)$, with $N \geq N_0$, is reasonably approximated by (22.58). Moreover, let Z_0 denote the sum

$$Z_0 \equiv \sum_{N=0}^{N_0-1} p_{24}(N)\, Z(M^2(N)). \tag{22.97}$$

This definition allows us to rewrite $Z_{\rm str}$ in (22.96) as

$$Z_{\rm str} = Z_0 + \sum_{N=N_0}^{\infty} p_{24}(N)\, Z(M^2(N)). \tag{22.98}$$

It is quite difficult to calculate Z_0 accurately. Our strategy will be to work in a regime where Z_0 is negligible compared to the second term on the right-hand side of (22.98). It will become clear below that this happens when the temperature T approaches the Hagedorn temperature. A few facts about Z_0 should be noted. The contribution from the tachyon states is problematic: equation (22.82) tells us that Z is a complex number if $m^2 < 0$. By neglecting Z_0 we are ignoring the tachyon instability in the theory. At any rate, Z_0 is a finite number for any value of the temperature, and its contribution to $Z_{\rm str}$ will be negligible when the second term on the right-hand side of (22.98) becomes very large.

To proceed further we approximate the sum in (22.98) by an integral. For $N \geq N_0$ we view $p_{24}(N)$ as the continuous function of N defined by the approximate relation (22.58). We then write

$$Z_{\rm str} \simeq Z_0 + \int_{N_0}^{\infty} dN\; p_{24}(N)\, Z(M^2(N)). \tag{22.99}$$

It is customary to define a density of states $\rho(M)$ as a function of the mass M and to use the mass as the variable of integration. This is done using the relation

$$p_{24}(N) dN = \rho(M) dM. \tag{22.100}$$

We express the left-hand side in terms of mass by using $\alpha' M^2 \simeq N$:

$$dN = 2\alpha' M dM = 2(\sqrt{\alpha'}M)\, d(\sqrt{\alpha'}M). \tag{22.101}$$

Moreover, using (22.58) and (22.69) we find

$$p_{24}(N) \simeq \frac{1}{\sqrt{2}} (\sqrt{\alpha'}M)^{-27/2} \exp(\beta_{\rm H} M). \tag{22.102}$$

Substituting these two equations back into (22.100) gives

$$\rho(M) dM = \sqrt{2}\, (\sqrt{\alpha'}M)^{-25/2}\, \exp(\beta_{\rm H} M)\, d(\sqrt{\alpha'}M). \tag{22.103}$$

Note incidentally that

$$\rho(M) \sim M^{-25/2} \exp(\beta_{\rm H} M), \tag{22.104}$$

which shows that the exponential growth in the density of states is controlled by the Hagedorn temperature. As we will see shortly, the partition function does not converge for temperatures higher than the Hagedorn temperature.

With (22.103) and (22.100), the partition function in (22.99) becomes

$$Z_{\rm str} \simeq Z_0 + \sqrt{2} \int_{M_0}^{\infty} (\sqrt{\alpha'} M)^{-25/2} \exp(\beta_{\rm H} M)\, Z(M^2)\, d(\sqrt{\alpha'} M), \tag{22.105}$$

where $\alpha' M_0^2 = N_0$. It only remains to write the particle partition function (22.89) in terms of M and $kT_{\rm H}$. With the help of

$$\frac{M}{2\pi\beta} = 2(\sqrt{\alpha'} M)\, kT\, kT_{\rm H}\,, \quad \beta M = 4\pi (\sqrt{\alpha'} M)\, \frac{T_{\rm H}}{T}, \tag{22.106}$$

we find

$$Z(M^2) \simeq 2^{25/2}\, V\, (kT\, kT_{\rm H})^{25/2}\, (\sqrt{\alpha'} M)^{25/2} \exp\Big(-4\pi\, \sqrt{\alpha'} M\, \frac{T_{\rm H}}{T}\Big). \tag{22.107}$$

Substituting this result into (22.105), the string partition function becomes

$$Z_{\rm str} \simeq Z_0 + 2^{13}\, V\, (kT\, kT_{\rm H})^{25/2} \int_{M_0}^{\infty} d(\sqrt{\alpha'} M) \exp\Big(-4\pi \sqrt{\alpha'} M \Big[\frac{T_{\rm H}}{T} - 1\Big]\Big). \tag{22.108}$$

Notice that the powers of M in the integrand cancelled out. Setting $x = \sqrt{\alpha'} M$, the above expression turns into

$$Z_{\rm str} \simeq Z_0 + 2^{13}\, V\, (kT\, kT_{\rm H})^{25/2} \int_{\sqrt{N_0}}^{\infty} dx\, \exp\Big(-4\pi x \Big[\frac{T_{\rm H}}{T} - 1\Big]\Big). \tag{22.109}$$

The integral only converges for $T < T_{\rm H}$, where we have

$$Z_{\rm str} \simeq Z_0 + \frac{2^{11}}{\pi}\, V\, (kT\, kT_{\rm H})^{25/2} \Big(\frac{T}{T_{\rm H} - T}\Big) \exp\Big(-4\pi \sqrt{N_0} \Big[\frac{T_{\rm H}}{T} - 1\Big]\Big). \tag{22.110}$$

In the limit when $T \to T_{\rm H}$ from below, the argument of the exponential goes to zero and as a result, the exponential goes to one. In addition, the factor which multiplies the exponential grows without limit and becomes much larger than Z_0. It follows that for T sufficiently close to $T_{\rm H}$, the partition function $Z_{\rm str}$ is well approximated by

$$Z_{\rm str} \simeq \frac{2^{11}}{\pi} V\, (kT_{\rm H})^{25} \Big(\frac{T_{\rm H}}{T_{\rm H} - T}\Big), \quad T \to T_{\rm H}. \tag{22.111}$$

In writing this formula we replaced T by $T_{\rm H}$ in all places where it is possible to do so. This is our final expression for the approximate partition function of a single open string that is enclosed in a box of volume V and is in thermal contact with a reservoir at a temperature T very close to $T_{\rm H}$.

We can use this result to calculate the average energy of a string near the Hagedorn temperature. In view of (22.6), we need only the β dependence of $\ln Z_{\rm str}$. Using $\beta \simeq \beta_{\rm H}$, the result is

$$\ln Z_{\rm str} \simeq -\ln(\beta - \beta_{\rm H}) + \cdots , \tag{22.112}$$

where the dots represent terms without β dependence. It now follows that the average energy $E_{\rm str}$ of the string is

$$E_{\rm str} = -\frac{\partial \ln Z_{\rm str}}{\partial \beta} \simeq \frac{1}{\beta - \beta_{\rm H}} \simeq kT_{\rm H}\Big(\frac{T_{\rm H}}{T_{\rm H} - T}\Big). \tag{22.113}$$

The energy $E_{\rm str}$ grows without bound as the temperature approaches the Hagedorn temperature.

String thermodynamics is not yet well understood. Some of the complications arise because in string theory the relation between the canonical and microcanonical ensembles is not familiar. In general, if we let $\Omega(E)dE$ be the number of states in the energy interval dE, the canonical partition function Z is defined as

$$Z(\beta) = \int_0^\infty dE\, e^{-\beta E}\, \Omega(E). \tag{22.114}$$

The microcanonical distribution $\Omega(E)$ determines an energy–temperature relation $E(T)$ via the relation

$$\beta = \frac{\partial \ln \Omega}{\partial E}, \tag{22.115}$$

as you recall from (22.2). In the canonical ensemble, this energy–temperature relation is roughly reproduced whenever the integral (22.114) is dominated by a saddle point. The saddle point arises at the maximum of the integrand, and it is determined by the condition

$$\frac{d}{dE}(e^{-\beta E}\,\Omega(E)) = \Big(-\beta\Omega(E) + \frac{\partial\Omega}{\partial E}\Big)e^{-\beta E} = 0. \tag{22.116}$$

As you can see, the vanishing condition reproduces (22.115). If the integral is dominated by this saddle point, the average energy $\bar{E}(T)$ calculated from Z turns out to be approximately equal to $E(T)$. In familiar systems $\Omega(E) \sim E^\gamma$, with $\gamma > 0$. In this case a saddle point exists, and the two ensembles give approximately equal results. On the other hand, in string theory $\Omega(E) \sim \exp(\beta_{\rm H} E)$, and the integrand is then proportional to $\exp([-\beta + \beta_{\rm H}]E)$, which is a function of energy that has no critical point. As a consequence, there is no guarantee that the two ensembles give the same results. It becomes important to decide physically which ensemble is relevant for each specific problem. Some refined computations using the microcanonical ensemble are examined in Problem 22.7. The results are surprising.

22.6 Black holes and entropy

A black hole is a gravitationally collapsed object that is formed when the mass of an object is increased while its size remains fixed, or when the size of an object is reduced while its mass is kept constant. Black holes were initially predicted by theoreticians before there were any observational data to support their existence, but it has now become an experimental fact that they exist in our universe. The presence of a supermassive black hole at the center of our galaxy has been established convincingly. Most likely, there are millions of black holes in every galaxy. They are the remnants of ordinary stars that were a few times more massive than the sun.

Black holes pose very significant theoretical challenges. In Einstein's theory of general relativity they appear as classical solutions which represent matter that has collapsed down to a point with infinite density: a singularity. Although dealing with classical singularities is already a theoretical challenge, the real puzzles of black holes arise at the quantum level. Quantum mechanically, black holes radiate energy. They also have thermodynamical temperature and entropy, but these properties are difficult to understand at the fundamental level of statistical mechanics, where they should be derived from a counting of degrees of freedom. String theory has had some impressive success in understanding the entropy of black holes. In this section we review basic features of black holes and use string theory to discuss the entropy of four-dimensional Schwarzschild black holes. In the following section we will examine a particular five-dimensional black hole for which the entropy can be calculated exactly in string theory.

The simplest black holes are Schwarzschild black holes. These black holes are spherically symmetric static solutions of Einstein's equations that represent the gravitational field of a point mass M. For such a black hole, the point singularity is separated from the outside world by what is known as an *event horizon*. This is a mathematical two-sphere centered at the singularity, whose radius R is called the Schwarzschild radius, or simply the radius of the black hole. If any object ventures inside the event horizon it will fall irrevocably into the singularity. Classically, nothing can escape from the region enclosed by the event horizon. The value R of the Schwarzschild radius can be estimated by assuming that the total energy of any particle at the horizon is equal to zero. For a particle of mass m, this energy includes the rest energy mc^2 and the gravitational potential energy $-GMm/R$. Setting the sum of these two equal to zero, we find

$$mc^2 - \frac{GMm}{R} = 0 \longrightarrow R \simeq \frac{GM}{c^2}. \tag{22.117}$$

The exact radius is calculated in general relativity and the answer is

$$R = \frac{2GM}{c^2}. \tag{22.118}$$

Physicists often speak of the Schwarzschild radius associated with a mass, meaning by this the radius of a black hole carrying such mass. The Schwarzschild radius of the sun is about three kilometers. The Schwarzschild radius of the earth is about one centimeter. The

Schwarzschild radius of a billion-ton asteroid is of the order of 10^{-15} m. It is possible to use the Newtonian laws of gravitation to estimate the gravitational field at the horizon:

$$|\vec{g}| = \frac{GM}{R^2} = \frac{c^4}{4GM}. \tag{22.119}$$

This gravitational field becomes small for very massive black holes. A massive object is a black hole if it is enclosed by a sphere that has its Schwarzschild radius.

Quick calculation 22.4 Show that a spherical object of uniform mass density ρ is a black hole if its radius is *larger* than $c/\sqrt{8\pi G\rho/3}$.

If we believe the second law of thermodynamics, then the existence of black holes leads to some surprising conclusions. Assume that a certain amount of hot gas falls into a black hole, so that the black hole now has a slightly higher mass. Since the total entropy of the system consisting of the gas and the black hole cannot decrease, the black hole must have acquired at least as much entropy as was carried by the gas. We are thus led to believe that black holes have entropy. You know that a system has entropy when there are many microscopic states of the system that are consistent with its macroscopic properties. On the other hand, if the black hole is simply a point mass singularity, it is hard to see what are the microstates that give rise to the entropy.

Black holes emit thermal radiation with a well defined temperature: a Hawking temperature $\bar{T}_{\rm H}$ that is proportional to the gravitational field at the horizon and, consequently, inversely proportional to the mass of the hole (do not confuse $\bar{T}_{\rm H}$ with the Hagedorn temperature $T_{\rm H}$). This proportionality is reasonable since the radiation from the black hole emerges from the near-horizon region, and is controlled by the intensity of gravity. In natural units the Hawking temperature of a black hole of mass M turns out to be $k\bar{T}_{\rm H} = 1/(8\pi M)$, so by inserting back the factors of $\hbar, c$, and G (Problem 3.7) we find

$$k\bar{T}_{\rm H} = \frac{\hbar c^3}{8\pi GM}. \tag{22.120}$$

This equation allows us to calculate the Bekenstein entropy $S_{\rm B}$ of the black hole. Using $E = Mc^2$ for the energy of the black hole and the first law of thermodynamics $dE = \bar{T}_{\rm H} dS_{\rm B}$, we write

$$dE = c^2 dM = \bar{T}_{\rm H} dS_{\rm B} = \frac{\hbar c^3}{8\pi GM}\,\frac{1}{k}\,dS_{\rm B}. \tag{22.121}$$

A little rearrangement yields

$$\frac{1}{k}\,dS_{\rm B} = \frac{4\pi G}{\hbar c} dM^2. \tag{22.122}$$

Integrating this equation, and assuming that the entropy of a zero-mass black hole is zero, we find

$$\frac{S_{\rm B}}{k} = \frac{4\pi G}{\hbar c}\,M^2. \tag{22.123}$$

The entropy of the black hole is proportional to the *square* of its mass. A useful alternative expression for the entropy uses the area A of the event horizon. With $A = 4\pi R^2$ and R given by (22.118), one readily obtains

$$\frac{S_{\rm B}}{k} = \frac{1}{4}\frac{c^3}{\hbar G} A = \frac{A}{4\ell_{\rm P}^2}, \tag{22.124}$$

where $\ell_{\rm P}$ is the Planck length. The right-hand side in this equation has a simple interpretation: the Bekenstein entropy $S_{\rm B}$ is one-fourth of the area of the horizon expressed in units of Planck-length squared. Given that $\ell_{\rm P}^2$ is a remarkably small area, the entropy of any astrophysical-size black hole is extremely large. The entropy of a black hole is roughly reproduced if one imagines having a degree of freedom with a finite number of states for each horizon element of area $\ell_{\rm P}^2$. String theory provides candidate degrees of freedom for black holes, but they do not relate directly to the horizon area.

Quick calculation 22.5 Show that a photon of energy $k\bar{T}_{\rm H}$ has a wavelength which is approximately 80 times the radius of the black hole.

In string theory, we attempt to relate a stationary Schwarzschild black hole to a string with a high degree of excitation but zero momentum. In the microcanonical ensemble, a string state with energy E has an entropy (22.67). This is true both for open and for closed strings (see (22.74)). Identifying $E = M$ and working henceforth with $\hbar = c = 1$, we have

$$\frac{S_{\rm str}}{k} = 4\pi\sqrt{\alpha'}\, M, \tag{22.125}$$

where we have added the subscript "str" to refer to the entropy of the string. This result should be compared to the black hole entropy (22.123):

$$\frac{S_{\rm B}}{k} = 4\pi G M^2. \tag{22.126}$$

The disagreement appears to be clear: the entropy of a black hole goes like the mass-squared, while the entropy of a string goes like the mass. We will soon show, however, that the apparent disagreement was to be expected. When they are properly understood, these equations display a surprising level of agreement. The linear dependence of the string entropy on the mass M of the string is not surprising. Entropy is an extensive quantity, and for a string the mass M is roughly proportional to its length L. The black hole entropy, on the other hand, exhibits a surprising feature: it is not proportional to the volume enclosed by the event horizon, but rather, to the area enclosed by the horizon. This failure of extensivity is a feature of gravitational physics.

Before considering the relation between equations (22.125) and (22.126), let us give a heuristic derivation of the string entropy. For this, we consider a string of mass M and estimate its length L to be roughly given by

$$M \sim T_0 L \sim \frac{1}{\alpha'} L, \tag{22.127}$$

where $T_0 \sim 1/\alpha'$ is the string tension. We now imagine a string built by joining together bits of string, each of which is of length $\ell_s = \sqrt{\alpha'}$. Each string bit can point in any of n possible directions. The number n may be equal to the number of spatial dimensions, but since our arguments are rough, we will not be specific. Since the number of string bits is $L/\sqrt{\alpha'}$, the number of ways Ω that we can build this string is roughly

$$\Omega \sim n^{L/\sqrt{\alpha'}} \sim n^{M\sqrt{\alpha'}} \sim e^{M\sqrt{\alpha'}\ln n}. \tag{22.128}$$

The entropy of the string is obtained by taking the logarithm of Ω:

$$\frac{S_{\rm str}}{k} \sim M\sqrt{\alpha'} \sim M\,\ell_s, \tag{22.129}$$

where we discarded the $\ln n$ factor, in keeping with the accuracy of the estimate. This result is consistent with the expression given in (22.125).

Equations (22.125) and (22.126) disagree because the black hole entropy $S_{\rm B}$ was calculated in a regime where interactions are necessary, while the string entropy $S_{\rm str}$ was calculated for free strings. We should not have expected agreement, unless for some reason interactions did not affect the calculation of the entropy of strings.

Interactions are necessary in the black hole entropy calculation because Newton's constant G vanishes if the string coupling constant g is set to zero. Indeed, we recall (13.83), which states that

$$G \sim g^2\,\alpha' = g^2\ell_s^2. \tag{22.130}$$

The black hole entropy and the black hole radius are then given by

$$\begin{aligned} \frac{S_{\rm B}}{k} &\sim GM^2 \sim g^2\ell_s^2 M^2, \\ R \ &\sim GM \ \ \sim g^2\ell_s^2 M. \end{aligned} \tag{22.131}$$

While they incorporate the string coupling dependence via Newton's constant, the above results use classical general relativity, where, for example, the concept of a horizon makes sense. We are allowed to neglect string theory corrections to general relativity as long as black holes are much larger than the string length.

Consider now a large black hole with Bekenstein entropy S_0, mass M_0, and radius $R_0 \gg \ell_s$. Fix also the string coupling to some finite value g_0. Equations (22.6) then give us

$$\begin{aligned} \frac{S_0}{k} &\sim g_0^2\,\ell_s^2\,M_0^2\,, \\ R_0 \ &\sim g_0^2\,\ell_s^2\,M_0\,. \end{aligned} \tag{22.132}$$

Since the calculation of the string entropy is valid for zero, and possibly small, string coupling, imagine now the process of dialing down the value of the string coupling. This is done by changing the expectation value of the dilaton, as explained in Section 13.4. It is reasonable to assume that this process can be carried out reversibly, so we can expect the black hole entropy to remain unchanged. On the other hand, as we dial down the string coupling g, the mass of the black hole must increase like $1/g$ to keep the entropy constant in (22.6). The mass is not increasing, however, if it is measured in units of Planck mass,

since $G \sim 1/m_{\rm P}^2$. The radius R of the black hole decreases, as follows from the second relation in (22.13) bearing in mind that $M \sim 1/g$.

Let g_*, R_*, and M_* denote the final values of the string coupling, black hole radius, and black hole mass, respectively. The constancy of the entropy and the formula for the radius give

$$\begin{aligned} \frac{S_0}{k} &\sim g_0^2\, \ell_s^2\, M_0^2 = g_*^2\, \ell_s^2 M_*^2\,, \\ R_* &\sim g_*^2\, \ell_s^2\, M_*. \end{aligned} \tag{22.133}$$

We do not expect these results to hold when the black hole becomes smaller than the string length, so let us fix $R_* = \ell_s$ as the minimum radius for which equations (22.133) can be trusted. The condition $R_* = \ell_s$ tells us that

$$g_*^2\, \ell_s^2\, M_* \sim \ell_s \;\longrightarrow\; M_* \sim \frac{1}{g_*^2\, \ell_s}. \tag{22.134}$$

Back in the expression for the entropy S_0 we find

$$\frac{S_0}{k} \sim \frac{1}{g_*^2}. \tag{22.135}$$

The coupling g_* is clearly very small since S_0 was assumed to be very large. At such weak coupling we can reasonably trust the free string theory expression (22.129) for the entropy. Since the black hole we are comparing with has mass M_*, we consider a string of mass M_*. The entropy is then given by

$$\frac{S_{\rm str}}{k} \sim M_*\, \ell_s \sim \Big(\frac{1}{g_*^2 \ell_s}\Big)\, \ell_s \sim \frac{1}{g_*^2}, \tag{22.136}$$

where we made use of (22.134). Comparing with (22.135), we see that $S_{\rm str} \sim S_0$. This agreement is evidence for the hypothesis that a Schwarzschild black hole is the strong coupling version of a string with a very high degree of excitation. It is far from a proof, however. As you have seen, we have only written approximate relations, and we have made a series of assumptions about the ranges of validity of certain results. A proof remains to be found at this time. Nevertheless, there is additional circumstantial evidence that this picture is at least roughly correct. It is possible to estimate the "size" of a string using the picture of string bits and assuming that the string is a random walk. One can then show that for any fixed coupling g there is a mass beyond which any excited string state is smaller than its Schwarzschild radius (Problem 22.9). This suggests that very heavy string states will form black holes.

22.7 Counting states of a black hole

Our computation of the entropy of strings can be done in the limit when we neglect the effects of interactions. Since a black hole can only exist once interactions are turned on, an

exact computation of the entropy of a black hole in string theory requires that the counting of states done with string coupling $g = 0$ remains valid when $g \neq 0$.

For the Schwarzschild black holes considered in the previous section this does not happen. As a result, we could only confirm qualitative agreement over a narrow range of couplings for which the gravity computation and the free string theory computation could both hold. In this section we wish to consider a particular five-dimensional black hole that appears in superstring theory. For this black hole, as we will explain below, the counting of states at zero string coupling will remain valid when the coupling becomes nonzero. It is the simplest known black hole with this property. Four-dimensional black holes with the same property are known, but they are are slightly more complicated. This is why we focus here on the five-dimensional black hole.

The remarkable property is due to supersymmetry. As long as supersymmetry is present, certain quantities can be calculated at zero coupling, and the results remain valid for all values of the coupling. Superstrings living in ten-dimensional Minkowski spacetime have supersymmetry. It is a challenge to compactify spacetime and preserve supersymmetry, but this happens if we curl up dimensions into circles. If we now include a black hole in the spacetime, supersymmetry can be lost. The black hole we are interested in is special: even when it is included some supersymmetry survives.

The starting point is ten-dimensional type IIB closed superstring theory. One can search for black hole solutions in the regime where the string theory is well approximated by a field theory of gravity, Kalb–Ramond fields, and other fields, including fermions. Such a theory is called type IIB supergravity. We curl up five of the spatial dimensions into circles. Let us denote these dimensions by x^5, x^6, x^7, x^8, and x^9. The black hole is a spherically symmetric configuration in the uncompactified effective spacetime M^5 defined by the coordinates x^0, x^1, x^2, x^3, and x^4. We cannot discuss here the full construction of the black hole, so we will simply summarize the results that are obtained.

(1) The black hole carries three different electric charges with respect to three Maxwell-like gauge fields that live on M^5. These charges are denoted by the integers

$$Q_1, \ Q_5, \text{ and } N. \tag{22.137}$$

A specific black hole is obtained by choosing these three integers.

(2) The black hole is *extremal*: it has the minimal mass that is compatible with its charges. It does not radiate, since radiation would reduce its mass without the necessary change of charge. The black hole has zero temperature. In addition, its presence preserves a large part of the original supersymmetry of the IIB theory in ten-dimensional Minkowski spacetime.

(3) The black hole horizon is a three-sphere with finite volume $A_{\rm H}$. The thermodynamical black hole entropy $S_{\rm bh}$ is calculated using the five-dimensional analog of (22.124):

$$\frac{S_{\rm bh}}{k} = \frac{A_{\rm H}}{4G^{(5)}} = 2\pi\sqrt{NQ_1Q_5}. \tag{22.138}$$

Here $G^{(5)}$ is the five-dimensional Newton constant and we have set $\hbar = c = 1$. Interestingly, the entropy only depends on the charges carried by the black hole and not

on other parameters, such as the string coupling or the size of the circles used for the compactification.

The goal is to use string theory to reproduce the entropy (22.138) by counting states. String theory must explain how this black hole can be constructed in many possible ways. We know how to count states in string theory when there are no interactions. This time, however, the black hole respects supersymmetry and this guarantees that the zero coupling counting continues to hold for nonzero coupling.

At zero coupling the black hole is constructed using type IIB superstring theory, with the five coordinates $x^5, \ldots, x^9$ curled up into circles. The charges Q_1 and Q_5 are generated by wrapping a number Q_1 of D1-branes around the circle x^5 and a number Q_5 of D5-branes around the five circles. These charges arise by the mechanism discussed in Section 16.4. Since a D5-brane has five spatial dimensions, the D5-branes wrap completely around the compact extra dimensions. How does this look to the five-dimensional observer in M^5? Since all spatial directions along M^5 are Dirichlet for the D5-branes, the D5-branes have fixed spatial coordinates on M^5. They appear, for all times, as a collection of static points. The same is true for the D1-branes. In the configuration we are trying to build, we require that all these points coincide. Thus all D-branes are coincident, and they are seen by the observer as a single point in space. This point is the center of the black hole that forms when the coupling is turned on. So far, this configuration of D-branes cannot be built in different ways that preserve supersymmetry. A few discrete choices are possible; we can choose, for example, another coordinate to wrap all of the D1-branes. But any number of choices that happens to be independent of the charges cannot give us the correct entropy. So, where does the entropy come from?

Recall that the macroscopic black hole had an additional charge N. What does this correspond to in the brane construction? It is a momentum quantum number. The momentum around the circle x^5 must equal

$$p^5 = \frac{N}{R}, \tag{22.139}$$

where R is the radius of the circle. This momentum cannot be carried by the D-branes since they are translationally invariant along the x^5 direction. The momentum is carried by open strings attached to the D-branes! We can now see how it is possible to get many states: there are many kinds of strings stretching between the Q_1 D1-branes and the Q_5 D5-branes. We have (1,1) strings going from D1-branes to D1-branes. We have (5,5) strings going from D5-branes to D5-branes. Finally, we have (1,5) and (5,1) strings, going from D1-branes to D5-branes and vice versa, respectively. Moreover, the total momentum quantum number N can be split between many open strings. Supersymmetry, however, makes one extra demand: all of the open strings must carry momentum in the same direction along x^5.

To proceed further, we need some known facts about the combined system of coincident D1- and D5-branes.

(1) The D1/D5-brane system is a bound state. Open strings of type (1,1) and (5,5) become massive and do not become excited in the configuration we are interested in. These strings can be dropped from the counting.

(2) The total number of ground states of a (1,5) string and the oppositely oriented (5,1) string is eight: four bosonic ground states and four fermionic ground states.
(3) The Q_1 D1-branes may join to form a single D1-brane wrapped Q_1 times around the circle. Similarly, the Q_5 D5-branes can join to form a single D5-brane wrapped Q_5 times around the full compact space. If this happens, the charges are not changed.

Bearing this information in mind, we see that the momentum number N must be split among open strings that stretch between D1-branes and D5-branes. We need a partition of N, but of which kind? Let us assume, for the time being, that $N \gg Q_1 Q_5$ and do a preliminary counting that will work but is not generally valid.

We have to partition N, and, for each element of each partition, we have to tell what kind of state is carrying the momentum quantum number. There are $Q_1 Q_5$ ways of picking a D1-brane and a D5-brane. Moreover, there are four additional ways to pick a bosonic excitation or, alternatively, four ways to pick a fermionic excitation (see (2) above). As a result, we have $b = 4Q_1 Q_5$ bosonic labels and $f = 4Q_1 Q_5$ fermionic labels. Making use of (22.62), the entropy is then

$$\frac{S_{\rm str}}{k} = \ln P\big(N;\, 4Q_1 Q_5,\, 4Q_1 Q_5\big) \sim 2\pi \sqrt{\frac{N}{6}(4Q_1 Q_5)\frac{3}{2}} = 2\pi\sqrt{N Q_1 Q_5}, \qquad (22.140)$$

in perfect agreement with (22.138). This is very nice, but it is not general enough. The restriction $N \gg Q_1 Q_5$ is needed because in (22.62) N must be much larger than both b and f. It can be shown that if N, Q_1, and Q_5 all grow large simultaneously then $\ln P$ fails to give the expected entropy. This means that we have not quite yet identified the general counting that gives the entropy.

The clue is given in item (3) of the list above. Imagine the D1-brane wrapped Q_1 times around the circle x^5. Consider then a (1,1) string moving along the D1-brane. How is the momentum of the string quantized? For such a string the circle has effectively become Q_1 times longer: $(2\pi R)Q_1$ is the distance the string must travel to return to its original starting point on the D1-brane. Accordingly, the string momentum is quantized in units of $1/(Q_1 R)$. This is true with one proviso. The individual open strings can have their momentum quantized with this finer unit, but the total momentum of all the open strings must still be quantized in units of $1/R$. This is because the system of the D1-brane and the attached open strings must be invariant under a translation by $2\pi R$ along the circle. As a result, the total momentum of the system must be quantized in units of $1/R$. Since the D1-brane has no momentum, the claim follows.

We must focus, however, on the strings stretching between D1-branes and D5-branes. Imagine now that the D5-branes are also wrapped. For simplicity, assume that Q_1 and Q_5 are relatively prime (we will relax this assumption shortly). Consider now a (1,5) string. How many times must it go around the x^5 circle so that *both* of its endpoints return to their original positions? After Q_1 turns the first endpoint does, but not the second. After Q_5 turns the second endpoint returns to its starting point, but the first does not. It takes $Q_1 Q_5$ turns to have both endpoints return to their original positions on the respective branes. As a result, the momentum of (1,5) and (5,1) strings is quantized with the even finer unit of $1/(Q_1 Q_5 R)$! This can be arranged to be approximately true even if Q_1 and Q_5 are

not relatively prime. Take, for example $Q_1 = Q_5 = 100$. We can take the D1-brane and split off one turn to get a system with $Q_1' = 99$ plus one extra D1-brane. Since Q_1' and Q_5 are relatively prime, the momentum of most open strings is then quantized in units of $1/(Q_1' Q_5 R)$, which is approximately equal to $1/(Q_1 Q_5 R)$. In general, for large Q_1 and Q_5 we can find relatively prime numbers $Q_1' < Q_1$ and $Q_5' < Q_5$ such that $Q_1' \sim Q_1$ and $Q_5' \sim Q_5$.

With this finer unit of quantization, the total momentum in (22.139) is suggestively written as

$$p^5 = \frac{N Q_1 Q_5}{Q_1 Q_5 R}. \tag{22.141}$$

This time we must partition the number $N Q_1 Q_5$. Since we just have one long D1-brane and one D5-brane, there is only one kind of string stretching between the branes. Therefore the labels on the elements of a partition are either four bosonic ones or four fermionic ones: $b = f = 4$. As a result, the entropy is given by

$$\frac{S_{\text{str}}}{k} = \ln P\,(N Q_1 Q_5;\ 4,\ 4) \sim 2\pi \sqrt{\frac{N Q_1 Q_5}{6}(4)\frac{3}{2}} = 2\pi\sqrt{N Q_1 Q_5}. \tag{22.142}$$

The agreement with the black hole entropy is now complete and holds generally.

The statistical mechanical derivation of black hole entropy is a significant accomplishment of string theory. Moreover, good progress has been made in extending the set of black holes for which the entropy is calculable. Still, much work remains to be done in string theory to understand black holes fully. Schwarzschild black holes are not under any precise control, and there are puzzles associated with the fate of the information that falls into a black hole.

String theory gives a clear picture of the zero-coupling degrees of freedom of a configuration that turns into a black hole for nonzero coupling. Moreover, we know that the counting continues to hold for nonzero coupling. We would like to know how these degrees of freedom look by the time the black hole is formed. Many mysteries remain.

Problems

Problem 22.1 Review of statistical mechanics.

(a) Prove equation (22.7). [Hint: the energy levels $E_\alpha(V)$ of the system depend on the volume. As the volume changes quasistatically, the change in mean energy is calculated using the equilibrium distribution of states. The change in mean energy can be interpreted as due to work against the pressure.]

(b) Prove equation (22.15). [Hint: consider the differential $d \ln Z(T, V)$.]

Problem 22.2 Fermionic violin string and counting unequal partitions.

Consider a system of simple harmonic oscillators with frequencies $\omega_0, 2\omega_0, \ldots$ identical to those of the bosonic violin string oscillators of Section 22.2. This time, however, each

occupation number n_ℓ can only take the value 0 or 1. Oscillators with this property are said to be fermionic oscillators.

(a) Calculate the free energy of such a string in the high temperature limit. The answer involves the sum
$$1 - \frac{1}{2^2} + \frac{1}{3^2} - \frac{1}{4^2} + \frac{1}{5^2} - \cdots,$$
which can be calculated using (22.38).

(b) Let $q(N)$ denote the number of partitions of N into *unequal* pieces. Use your result in (a) above to show that the large-N behavior of $\ln q(N)$ is given by (22.61).

(c) Now assume this string is relativistic, with the energy related to the mode number as in (22.65): $\sqrt{N} = \sqrt{\alpha'}E$. What is the Hagedorn temperature for such a string?

Problem 22.3 Generating functions for partitions.

A particularly simple infinite product provides a generating function for the partitions $p(n)$:
$$\prod_{n=1}^{\infty} \frac{1}{(1-x^n)} = \sum_{n=0}^{\infty} p(n) x^n \,.$$
Here $p(0) \equiv 1$. To evaluate the left-hand side, each factor is expanded as an infinite Taylor series around $x = 0$. Test this formula for $n \leq 4$, and explain (in words) why it works in general. Find a generating function for unequal partitions $q(n)$, and test it for low values of n.

Problem 22.4 Counting of generalized partitions.

Prove the formula (22.62) for the partitions $P(N; b, f)$ of N into ordinary integers with b labels and fermionic integers with f labels. Calling Z the partition function of ordinary oscillators and Z' the partition function of Problem 22.2, begin your derivation by explaining why the partition function Z_T for the composite system of bosonic and fermionic labeled oscillators is given by
$$Z_T = (Z)^b \, (Z')^f \,.$$

Problem 22.5 Open superstring Hagedorn temperature.

Consider the open superstring theory described in Section 14.1.

(a) Show that the total number of states (NS and R sectors) with number $N^\perp$ is $16P(N^\perp; 8,8)$. [Hint: one of the two sectors is easier to count; then use supersymmetry.]

(b) Following the method of Section 22.3, calculate the Hagedorn temperature for an open superstring. Show that it is a factor of $\sqrt{2}$ larger than the Hagedorn temperature of the bosonic string.

Problem 22.6 Partition function of the relativistic particle.

Evaluate exactly the partition function (22.83) for the relativistic point particle in terms of modified Bessel functions (and derivatives thereof), making use of the integral representation

$$K_\nu(z) = \frac{\sqrt{\pi}\left(\frac{1}{2}z\right)^\nu}{\Gamma\left(\nu+\frac{1}{2}\right)} \int_0^\infty e^{-z\cosh t}\,\sinh^{2\nu} t\,dt\,.$$

Use the asymptotic expansion

$$K_\nu(z) \sim e^{-z}\sqrt{\frac{\pi}{2z}}\Big[1+\frac{4\nu^2-1}{8z}+\cdots\Big],$$

which is valid for large z, to confirm our low temperature result in (22.89). Calculate the first nontrivial correction to this result.

Problem 22.7 Corrections to the temperature/energy relation in the microcanonical ensemble.

We found the Hagedorn temperature in the idealized string model by computing the entropy–energy relation in the high energy approximation where $\ln p_{24}(N) \sim 4\pi\sqrt{N}$. Use the more accurate expression for the partitions $p_{24}(N)$ as given in (22.58) to find the corrections to the temperature–energy relation. You will find the surprising result that as the energy goes to infinity the temperature goes to $T_{\rm H}$ from above! Plot $T(E)$, and calculate the (negative!) specific heat C in the high energy regime.

The above computations were done using $S = k\ln p_{24}(N)$ and the relation between E and N. In conventional systems with continuous energies one defines $S = k\ln\Omega(E)$, where $\Omega(E)dE$ is the number of states in the energy interval dE. Use the relation $\Omega(E)dE = p_{24}(N)dN$ to calculate $\Omega(E)$, and show that the resulting entropy $S(E)$ differs slightly from the one calculated before.

The negative specific heat obtained for the idealized string is not atypical. For open strings on a Dq-brane one can prove that

$$\Omega(E) \simeq E^{-\gamma}\exp(4\pi\sqrt{\alpha' E})\,, \quad \text{with} \quad \gamma = (25-q)/2.$$

Use the continuous energy formulation to calculate the specific heat C in the high energy regime. Give your answer in terms of γ, E, and $kT_{\rm H}$.

Problem 22.8 Long strings are entropically favored.

The general approximate formula for partitions $p_b(N)$ takes the form

$$p_b(N) \sim \beta N^{-\gamma}\exp(\delta\sqrt{N}),$$

where β, γ, and δ are positive, b-dependent constants. Consider an open bosonic string with large excitation number $N^\perp = N_0$ and energy E_0 related by $\alpha' E_0^2 \simeq N_0$. Assume zero momentum for all strings.

(a) Find the ratio of the number of states available to the string and the number of states available when the string breaks into two (distinguishable) strings, each carrying half the energy. Express your answer in terms of N_0 and the constants in the problem. What is the change in entropy $\Delta S/k$ during the process in which the two half-strings join to make the original string? Show that this change is positive when N_0 is sufficiently large. Which constant out of β, γ, and δ is the one responsible for this effect?

(b) Show that, in general, the combination of two open strings with large excitation numbers into a single open string is a process that increases the entropy.
(c) Since the above results are valid for large N_0, it is of some interest to test them for smaller N_0. How many times are the number of available states increased when a string with $N_0 = 9$ is formed from two strings, each of which has the same energy? What is the change of entropy $\Delta S/k$? Use the exact formula $\alpha' E^2 = N - 1$ for all cases. (A little help: $p_{24}(9) = 143\,184\,000$.)

Problem 22.9 Estimating the size of a string state.

We used the heuristic picture of a string made out of string bits to estimate correctly the entropy (22.129) of a string. We now want to use this picture to estimate the size of an open string state. Assume that each string bit can point randomly in any of d orthogonal directions. The string can then be viewed as a random walk with a number of steps equal to the number of bits.

(a) Use the random walk formula for the average value of the square of the displacement to show that the "size" $R_{\rm str}$ of a string of mass M is

$$R_{\rm str}(M) \sim M^{1/2}\,\ell_s^{3/2} \sim N^{1/4}\ell_s, \tag{1}$$

where N is the number eigenvalue associated with the mass M. Note that the size grows like the square root of the mass, while the length of the string grows like the mass. $R_{\rm str}$ is the size of the string at zero string coupling.
(b) Show that the (zero string coupling) size $R_{\rm str}$ of a string of mass M is smaller than the Schwarzschild radius of a mass M in a theory with coupling g, when $M > \bar{M}$ where

$$\bar{M} \sim \frac{1}{g^4\ell_s} \sim \frac{m_{\rm P}}{g^3}. \tag{2}$$

This result suggests that the sufficiently massive strings could form a black hole. Show that for $\bar{M}$ the radius of the hole is ℓ_s/g^2, so for couplings that are small, the size of the hole is too large to trust the string model. Show that the value of N for a zero coupling string of mass $\bar{M}$ is $N \sim 1/g^8$. Give a rough estimate of $\bar{M}$ in kg when $g \sim 0.01$.
(c) Consider a very large black hole of mass M and Schwarzschild radius R in a string theory at some finite coupling g. Assume now that the coupling is slowly dialed down to zero. Calculate the value of N that characterizes the zero coupling string that is obtained (write your answer in terms of M and $m_{\rm P}$). Show that the size $R_{\rm str}$ of this string is equal to R/g. How much bigger is the length of the string?

The mass of the black hole at the center of our galaxy is approximately 2.6 million solar masses. Estimate the value of N for the corresponding zero coupling string. (Answer: $N \sim 10^{177}$.)

The random walk model of string states applies to strings with little or no angular momentum (recall that the size of a rigidly rotating open string is proportional to the mass). In this model the size of the string is much smaller than its length. The effect of string interactions appears to reduce further the size of the strings.

23 Strong interactions and AdS/CFT

String theory offers a number of insights into the theory of strong interactions. The quantum states of a rotating open string have key properties of hadronic excitations. The energy of a stretched string matches quite well the potential energy of a separated quark–antiquark pair. More surprisingly, certain strongly interacting gauge theories are physically equivalent to closed string theories. The closed strings propagate on a space whose boundary is roughly the space where the gauge theory lives. The prime example of this equivalence is the AdS/CFT correspondence, which states that supersymmetric four-dimensional $SU(N)$ gauge theory is fully described by type IIB closed superstrings in a spacetime that includes the five-dimensional anti-de Sitter space AdS_5. We motivate this correspondence and examine in detail the geometry of anti-de Sitter space and related hyperbolic spaces. The correspondence suggests that properties of the recently discovered quark–gluon plasma are related to properties of black holes in anti-de Sitter space.

23.1 Introduction

String theory was discovered in the attempts to understand the dynamics of strongly interacting hadrons. It had been noted that the plot of the angular momentum J of hadronic excitations against their energy-squared falls roughly into lines $J = \alpha' E^2$ called Regge trajectories. String theory seemed to be a reasonable candidate for a theory of strong interactions because this relationship between J and E^2 emerges naturally from a rotating classical open string, as we discussed in Section 8.6. Quantization picks discrete values for the angular momentum and modifies the linear relation between J and E^2 by the addition of a constant. Both changes are needed for agreement with the data.

Despite these encouraging indications, the early attempts to use relativistic strings to describe hadrons faced many problems. Among them was the presence of unwanted massless vectors and massless tensors, precisely the particles needed to make string theory a candidate for a unified theory of physics. The string theory approach to strong interactions was abandoned and quantum chromodynamics, or QCD, was adopted. QCD postulates that the basic constituents of hadrons are quarks and gluons. QCD is a quantum field theory, in fact, an $SU(3)$ Yang–Mills theory.

In QCD, a meson is viewed as a pair of quarks held together by gluons. The string picture, however, remains a useful approximate description. In this picture, a meson is a pair of quarks held at the ends of an open string that represents a thin tube of color flux

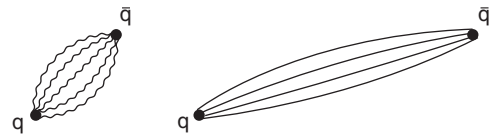

Fig. 23.1 **A meson as a quark–antiquark pair held together by color field lines. When the quark and antiquark are separated the color field lines form a thin flux tube that can be viewed as a string.**

lines (Figure 23.1). The confinement of quarks, that is, the fact that quarks are never seen in isolation, also has a simple explanation in the string picture. Since the tension of the string joining the quarks is independent of its length, it takes infinite energy to separate fully a pair of quarks. The potential energy between a quark and an antiquark has been the subject of much analytic and numerical work in QCD. As we will see, key features of this potential follow from a simple string model.

In the above string picture a meson is a string. When we discussed the Standard Model on a configuration of D6-branes, the quarks and leptons were the strings. There is no contradiction: a superstring may describe elementary particles and at the same time another string theory, perhaps an effective one, may describe the composite hadrons. In fact, it was a long-held belief that there should be some exact string theory description of QCD. While this string theory is not yet known, a surprising development has come tantalizingly close to achieving this objective.

As it turns out, certain strongly interacting gauge theories have an *exact* string picture, but the strings do not propagate in the spacetime where the strongly interacting theory lives. Roughly, the gauge theory lives on the boundary of the spacetime where the strings propagate. The equivalence of a gauge theory and a string theory was first exhibited in the AdS/CFT correspondence. In this correspondence, a maximally supersymmetric $SU(N)$ Yang–Mills theory in four-dimensional Minkowski spacetime is claimed to be equivalent to a type IIB closed superstring theory. The ten-dimensional spacetime in this superstring theory takes a particular form: five dimensions form a sphere S^5 and the other five dimensions form a noncompact *anti-de Sitter* spacetime, briefly denoted by AdS_5. One can view the Minkowski spacetime of the field theory as the boundary of the AdS_5 space. The maximally supersymmetric $SU(N)$ Yang–Mills theory has as much supersymmetry as a gauge theory can have. It is a conformal field theory (CFT), a type of theory that has no dimensionful parameters. In the AdS/CFT acronym, AdS stands for anti-de Sitter space and CFT stands for the gauge theory. After a discussion of the correspondence, we examine a variant that describes a "hot" gauge theory – the quark–gluon plasma – using strings moving on an AdS space that contains a black hole.

23.2 Mesons and quantum rotating strings

The classical linear relation $J = \alpha' M^2$ between angular momentum and mass-squared of a rotating open string suggests that the Regge trajectories of mesonic excitations could have

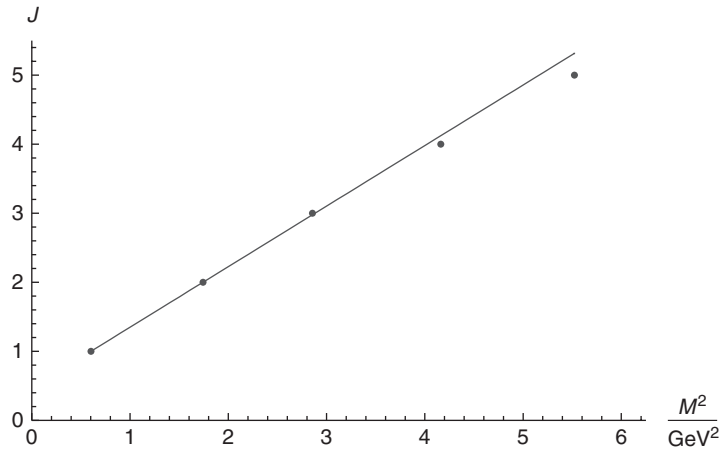

Fig. 23.2 **Frautschi–Regge trajectory for the rho-mesons $\rho(776)$, $a_2(1320)$, $\rho_3(1690)$, $a_4(2040)$, and $\rho_5(2350)$. The line was fit to go through the $\rho(776)$ and the $a_2(1320)$.**

a string theory explanation. Let us examine a particular meson and the Regge trajectory associated with it. We choose the rho-meson.

The rho-meson $\rho(776)$, of mass 776 MeV, is in fact a triplet of mesons (ρ^+, ρ^0, ρ^-), two of them charged and one uncharged. The rho-mesons are heavy analogs of the pions (π^+, π^0, π^-). Just like the pions, they are made of u and d quarks as well as antiquarks. As opposed to the pions, which are spin-zero combination of quarks, the rho-mesons have spin angular momentum S equal to one. With zero orbital angular momentum L, their total angular momentum is $J = 1$. The rho-mesons are unstable and decay mostly to a pair of pions, with lifetimes of about 10^{-23}s. The $\rho(776)$ belongs to a trajectory of mesons with $S = 1$ and higher values of L. Writing the masses (in MeV) inside parenthesis, we have the $a_2(1320)$ with $J = 2$, the $\rho_3(1690)$ with $J = 3$, the $a_4(2040)$ with $J = 4$, and the $\rho_5(2350)$ with $J = 5$. These five mesons are plotted in Figure 23.2.

To get a sense of the accuracy of the linear trajectory we fit a line through the lowest two mesons. We write

$$J = \alpha' M^2 + \beta' \,, \tag{23.1}$$

and determine the constants α' and β' so that $M^2 = (0.776\,\text{GeV})^2$ gives $J = 1$ and $M^2 = (1.320\,\text{GeV})^2$ gives $J = 2$. This gives

$$J = 0.877\,02\,(\text{GeV})^{-2}\,M^2 + 0.471\,88\,. \tag{23.2}$$

This is the line shown in Figure 23.2. Using this fit one predicts a $J = 3$ meson of mass 1699 MeV, a $J = 4$ meson of mass 2006 MeV and a $J = 5$ meson of mass 2272 MeV. The errors in the masses are really small: 0.5%, 1.7%, and 3.3%, respectively.

Having seen that (23.1) is needed to fit the data, let us see how the classical relation $J = \alpha' M^2$ is modified in the quantum theory. Let us assume that the open string rotates in

the (x^2, x^3) plane and has zero momentum in this plane. The relevant angular momentum operator $J = M^{23}$ is read from (12.147):

$$J = M^{23} = -i \sum_{n=1}^{\infty} \frac{1}{n} \left(\alpha_{-n}^{(2)} \alpha_n^{(3)} - \alpha_{-n}^{(3)} \alpha_n^{(2)} \right) . \tag{23.3}$$

Naturally, the terms on the above right-hand side mix the oscillators from the two spatial coordinates. It is useful to define new oscillators α_n and $\bar{\alpha}_n$:

$$\alpha_n \equiv \frac{1}{\sqrt{2}} (\alpha_n^{(2)} + i \alpha_n^{(3)}) , \qquad \bar{\alpha}_n \equiv \frac{1}{\sqrt{2}} (\alpha_n^{(2)} - i \alpha_n^{(3)}) . \tag{23.4}$$

Note that $(\alpha_n)^\dagger = \bar{\alpha}_{-n}$. A short calculation shows that the commutation relations for the new oscillators are

$$[\alpha_m, \bar{\alpha}_n] = m \, \delta_{m+n,0} , \qquad [\alpha_m, \alpha_n] = [\bar{\alpha}_m, \bar{\alpha}_n] = 0 . \tag{23.5}$$

Note that the first relation implies that $[\bar{\alpha}_m, \alpha_n] = m \delta_{m+n,0}$. In terms of the new oscillators the angular momentum operator is given by

$$J = \sum_{n=1}^{\infty} \frac{1}{n} \left\{ \alpha_{-n} \bar{\alpha}_n - \bar{\alpha}_{-n} \alpha_n \right\} . \tag{23.6}$$

The first term in the above sum counts the number of α oscillators in a state. The second term, up to a sign, counts the number of $\bar{\alpha}$ oscillators in a state. The convenience of the new basis is now apparent.

Quick calculation 23.1 Verify equation (23.6), and convince yourself of the claims that follow this equation.

Using the new oscillators we write states in the form

$$|\lambda\rangle = \ldots \prod_{k=1}^{\infty} (\alpha_{-k})^{\lambda_k} (\bar{\alpha}_{-k})^{\bar{\lambda}_k} |p^+, \vec{p}_T\rangle , \tag{23.7}$$

where $\lambda_k \geq 0$ and $\bar{\lambda}_k \geq 0$ are integers and the dots represent products of oscillators in directions other than x^2 and x^3. Acting on $|\lambda\rangle$

$$J \text{ has eigenvalue } \mathcal{J} = \sum_{k=1}^{\infty} (\lambda_k - \bar{\lambda}_k) . \tag{23.8}$$

Is it thus easy to build states with arbitrary values of the angular momentum.

Since we want to find the mass-squared values of states with definite angular momentum we consider the mass-squared relation (12.164) and write it as:

$$\alpha' M^2 + 1 = N^\perp = N_{23} + N' , \tag{23.9}$$

where

$$N_{23} = \sum_{n=1}^{\infty} \left(\alpha_{-n}^{(2)} \alpha_n^{(2)} + \alpha_{-n}^{(3)} \alpha_n^{(3)} \right) \tag{23.10}$$

denotes the contribution to $N^\perp$ from the x^2 and x^3 directions and N' denotes the contribution to $N^\perp$ from the other transverse directions. One readily finds that in terms of the new oscillators

$$N_{23} = \sum_{n=1}^{\infty} (\alpha_{-n}\bar{\alpha}_n + \bar{\alpha}_{-n}\alpha_n)\,, \tag{23.11}$$

and this implies that on the general states (23.7)

$$N_{23} \text{ has eigenvalue } \mathcal{N}_{23} = \sum_{k=1}^{\infty} k(\lambda_k + \bar{\lambda}_k)\,. \tag{23.12}$$

Since $\lambda_k, \bar{\lambda}_k \geq 0$ and $k \geq 1$, we have

$$\mathcal{N}_{23} = \sum_{k=1}^{\infty} k(\lambda_k + \bar{\lambda}_k) \geq \sum_{k=1}^{\infty} \lambda_k + \sum_{k=1}^{\infty} \bar{\lambda}_k \geq \left| \sum_{k=1}^{\infty} \lambda_k - \sum_{k=1}^{\infty} \bar{\lambda}_k \right| = |\mathcal{J}|\,, \tag{23.13}$$

where we used (23.8) and noted that for any two numbers $b_1 \geq 0$ and $b_2 \geq 0$, one has $b_1 + b_2 \geq |b_1 - b_2|$. The inequality we have obtained is

$$\mathcal{N}_{23} \geq |\mathcal{J}|\,. \tag{23.14}$$

Noting that the eigenvalues of N' are greater than or equal to zero, equation (23.9) gives the inequality $1 + \alpha'\mathcal{M}^2 \geq \mathcal{N}_{23}$, where $\mathcal{M}^2$ is the eigenvalue of M^2. Finally, combining this inequality with (23.14) we obtain

$$|\mathcal{J}| \leq 1 + \alpha'\mathcal{M}^2\,. \tag{23.15}$$

This inequality holds for arbitrary states in (23.7). It is the quantum version of the classical equation $J = \alpha' M^2$. For states that saturate the inequality we have $\mathcal{J} = \alpha'\mathcal{M}^2 + 1$, an equation of the type (23.1). While α' must be determined by fitting the data, β' is predicted. The value $\beta' = 1$, from bosonic string theory, does not fit the rho-meson trajectory well. While most Regge trajectories give rather similar values of α', the values of β' vary more widely.

Let us examine states that saturate the inequality (23.15). For a state of the form $(\alpha_{-1})^N|p^+, \vec{p}_T\rangle$ we have $\lambda_1 = N$, all other λs and $\bar{\lambda}$s = 0, and $N' = 0$. So

$$\mathcal{J} = N\,, \qquad 1 + \alpha'\mathcal{M}^2 = \mathcal{N}_{23} = 1 \cdot \lambda_1 = N = \mathcal{J}\,, \tag{23.16}$$

and the inequality is saturated. The state $(\alpha_{-1})^N|p^+, \vec{p}_T\rangle$ is said to belong to the Regge trajectory with maximal angular momentum per unit mass-squared.

The quantum states $|\psi_N\rangle \sim (\alpha_{-1})^N|p^+, \vec{p}_T\rangle$ resemble the classical rotating string: their angular momentum and mass-squared eigenvalues are related roughly like in the classical theory. In order to compute well defined expectation values the states must be normalized, and this requires delta functions for the momenta, as indicated in (12.171). In order to

keep the notation simple and avoid working with superpositions of states we just write $\langle p^+, \vec{p}_T | p^+, \vec{p}_T \rangle = 1$ instead of delta functions of zero argument.

Perhaps surprisingly, the expectation values $\langle \psi_N | X^I(\tau,\sigma) | \psi_N \rangle$ do not behave at all like the classical coordinates $X^I(\tau,\sigma)$ of a rotating string. We find

$$\langle \psi_N | X^I(\tau,\sigma) | \psi_N \rangle = \langle x_0^I \rangle + 2\alpha' p^I \tau \,, \tag{23.17}$$

because the oscillators in the mode expansion of $X^I(\tau,\sigma)$ cannot contribute: adding a single oscillator to a nonvanishing expectation value makes it vanish. According to the expectation value (23.17) the string propagates as if it were a point. One can see evidence that the string is extended by computing the expectation values of *squares* of the coordinates (see Problem 23.1).

There are, however, coherent states for which the expectation values of the coordinates follow the classically expected trajectory. Take, for example,

$$|\Psi\rangle = e^{\mathcal{A}} | p^+, \vec{0} \rangle \,, \quad \text{with} \quad \mathcal{A} = v\,(\bar{\alpha}_1 - \alpha_{-1}) \,, \tag{23.18}$$

where v is a real constant. Since $\mathcal{A}^\dagger = -\mathcal{A}$, we have $\langle \Psi | \Psi \rangle = 1$, up to the caveat discussed above. Let us first confirm that the expectation values of J and M^2 are the familiar ones. The expectation value of J is readily computed:

$$\langle J \rangle = \langle \Psi | J | \Psi \rangle = \langle \Psi | (\alpha_{-1}\bar{\alpha}_1 - \bar{\alpha}_{-1}\alpha_1) | \Psi \rangle \,, \tag{23.19}$$

since all other terms in (23.6) commute with $e^{\mathcal{A}}$ and annihilate the vacuum states to the right. In fact, α_1 also commutes with $\mathcal{A}$ and therefore the second term above does not contribute either. We then have

$$\begin{aligned} \langle J \rangle &= \langle \Psi | \alpha_{-1}\bar{\alpha}_1\, e^{\mathcal{A}} | p^+, \vec{0} \rangle = \langle \Psi | \alpha_{-1}\, [\bar{\alpha}_1 \,, e^{\mathcal{A}}\,] | p^+, \vec{0} \rangle \\ &= \langle \Psi | \alpha_{-1}\, [\bar{\alpha}_1 \,, \mathcal{A}\,] e^{\mathcal{A}} | p^+, \vec{0} \rangle = -v \langle \Psi | \alpha_{-1}\, e^{\mathcal{A}} | p^+, \vec{0} \rangle \,. \end{aligned} \tag{23.20}$$

Given that $\langle \Psi |\, e^{\mathcal{A}} \alpha_{-1} | p^+, \vec{0} \rangle = \langle p^+, \vec{0} | \alpha_{-1} | p^+, \vec{0} \rangle = 0$ we have

$$\langle J \rangle = -v \langle \Psi |\, [\alpha_{-1} \,, e^{\mathcal{A}}]\, | p^+, \vec{0} \rangle = -v \langle \Psi | [\alpha_{-1} \,, \mathcal{A}] e^{\mathcal{A}} | p^+, \vec{0} \rangle = v^2 \,. \tag{23.21}$$

Since the computation of N_{23} involves exactly the same term $\alpha_{-1}\bar{\alpha}_1$ used for the computation of $\langle J \rangle$, we have, all in all:

$$\langle J \rangle = \langle N_{23} \rangle = v^2 \,, \quad \langle \alpha' M^2 \rangle = v^2 - 1 \quad \to \quad \langle J \rangle = \langle (\alpha' M^2 + 1) \rangle \,, \tag{23.22}$$

as expected for a rotating string. Let us now consider the expectation values of the coordinates. A short computation using the oscillator expansion (12.66) gives

$$\langle (X^2 + i X^3)(\tau,\sigma) \rangle = \langle x_0^2 + i x_0^3 \rangle + 2i\sqrt{\alpha'} \langle (\alpha_1 e^{-i\tau} - \alpha_{-1} e^{i\tau}) \rangle \cos\sigma \,. \tag{23.23}$$

Noting that $\langle \alpha_1 \rangle = 0$ and $\langle \alpha_{-1} \rangle = -v$ and assuming $\langle x_0^2 \rangle = \langle x_0^3 \rangle = 0$, we find

$$\langle (X^2 + i X^3)(\tau,\sigma) \rangle = 2iv\sqrt{\alpha'} e^{i\tau} \cos\sigma \,. \tag{23.24}$$

Separating real and imaginary parts we indeed obtain the classical limit of a rotating string:

$$\begin{aligned}\langle X^2(\tau,\sigma)\rangle &= -2v\sqrt{\alpha'}\,\sin\tau\,\cos\sigma\,,\\ \langle X^3(\tau,\sigma)\rangle &= \;\;2v\sqrt{\alpha'}\,\cos\tau\,\cos\sigma\,.\end{aligned}\tag{23.25}$$

The parameter v can be related to familiar constants of the motion. The length ℓ of the above string is $\ell = 4v\sqrt{\alpha'}$. Moreover, since spatial momenta are zero, $M^2 = 2p^+p^- = 2(p^+)^2$. From this and (23.22) we find

$$p^+ = \frac{1}{\sqrt{2\alpha'}}\sqrt{v^2-1}\,.\tag{23.26}$$

In the classical limit v is large and

$$p^+ \simeq \frac{v}{\sqrt{2\alpha'}} = \frac{\ell}{4\sqrt{2}\,\alpha'}\,.\tag{23.27}$$

This is the classical relation between p^+ and the length of a rotating string. Indeed, using (7.60), which states that the energy E of such string is $\frac{\pi}{2}T_0\ell$,

$$p^+ = \frac{1}{\sqrt{2}}\,E = \frac{1}{\sqrt{2}}\,\frac{\pi}{2}\,T_0\ell = \frac{1}{\sqrt{2}}\,\frac{\pi}{2}\,\frac{1}{2\pi\alpha'}\,\ell = \frac{\ell}{4\sqrt{2}\alpha'}\,.\tag{23.28}$$

We have thus verified that the expectation values of the coordinates describe a rotating string of the correct size. The use of coherent states to describe quantum states with semiclassical limits is familiar from the simple harmonic oscillator, as reviewed in Problem 23.2.

23.3 The energy of a stretched effective string

The classical potential energy of a static stretched string is equal to the tension T_0 times its length ℓ. If we imagine a quark and an antiquark fixed at the ends of the string, this energy would represent the quark–antiquark potential $V(\ell)$ evaluated at a separation ℓ. The classical limit suggests that for large r one has $V(r) \sim T_0 r$.

How do we use quantum string theory to obtain information about $V(r)$? We look into the calculation of the mass-squared of a string stretched between two parallel D-branes. The D-branes are not good representations for the quarks, but they certainly hold the string stretched, and we are after the energy of the string, which we assume defines the potential $V(r)$. For D-branes separated a distance L we obtained (14.51), that written in terms of the string tension and an arbitrary spacetime dimension D reads:

$$M^2 = \big(T_0 L\big)^2 + \frac{1}{\alpha'}\Big(N^\perp - \frac{1}{24}(D-2)\Big)\,.\tag{23.29}$$

In order to represent a string without excitations we assume $N^\perp = 0$. Moreover, since the string is static we can identify M^2 with the potential energy squared. We thus write, using r for the length of the string,

$$\big(V(r)\big)^2 = (T_0\, r)^2 - \frac{1}{\alpha'}\frac{(D-2)}{24}\,. \tag{23.30}$$

At this stage we reach a significant complication. We would like to determine the quark–antiquark potential in *four* spacetime dimensions. We are tempted to set $D = 4$ in the above equation, but the quantum consistency of the Nambu–Goto action requires $D = 26$. We cannot justify using this equation for other values of D.

There is evidence, however, that a variant of the Nambu–Goto action exists that can be formulated with both Lorentz and translational symmetry in spacetimes of arbitrary dimensionality D. The theory is complicated and the action contains an infinite number of additional terms. The quantization of a stretched string in this theory gives a potential that roughly agrees with (23.30): it reproduces the first few terms in the large r expansion of $V(r)$.

Thus reassured, we proceed to explore (23.30) for arbitrary D. It is customary to introduce the Luscher coefficient

$$\gamma_D \equiv -\frac{\pi}{24}(D-2)\,, \tag{23.31}$$

whose values in $D = 4$ and $D = 3$ are

$$\gamma_4 = -\frac{\pi}{12} = -0.262, \quad \gamma_3 = -\frac{\pi}{24} = -0.1309\,. \tag{23.32}$$

We can then write (23.30) as

$$V(r) = \sqrt{(T_0 r)^2 + 2T_0\gamma_D}\,. \tag{23.33}$$

Expanding the above right-hand side for large r,

$$V(r) = T_0\, r + \gamma_D \cdot \frac{1}{r} + \mathcal{O}(1/r^3)\,. \tag{23.34}$$

The above gives the leading quantum correction to the classical potential $V(r) = T_0 r$ for large r. The tension $T(r)$ on the string is the magnitude of the force associated with the potential:

$$T(r) = \frac{\partial V}{\partial r} = T_0 - \gamma_D \cdot \frac{1}{r^2} + \mathcal{O}(1/r^4)\,. \tag{23.35}$$

The tension $T(r)$ is equal to T_0, up to a correction linear in $1/r^2$. Since $\gamma_D < 0$ for $D > 2$, $T(r) > T_0$ in three or four spacetime dimensions. The Luscher coefficient can be extracted through derivatives of the force:

$$C(r) \equiv \frac{1}{2} r^3 \frac{\partial T}{\partial r} = \gamma_D + \mathcal{O}(1/r^2)\,. \tag{23.36}$$

Note that $C(r) \sim \gamma_D$, with small corrections for large r. The quantity $C(r)$ can be determined numerically in the lattice approximation. The computations evaluate the quark–antiquark potential in the prototypical $SU(3)$ gauge theory. The results give strong evidence that as r is increased $C(r)$ approaches the predicted values (23.32) of γ_D both for

$D = 4$ and for $D = 3$. This indicates that the string picture captures not only the leading term $T_0 r$ of the quark–antiquark potential but also the first nontrivial correction!

23.4 A large-N limit of a gauge theory

We now begin our study of the ideas relevant to the AdS/CFT correspondence. This correspondence of gauge theory with gravity is clearest when the gauge theory has a large number of degrees of freedom. We will thus examine an $SU(N)$ gauge theory with a large value of N. This theory has a dimensionless coupling constant $g_{\rm YM}$ which controls the strength of the interactions between gauge bosons. More concretely, each time a gauge boson turns into two gauge bosons, or two gauge bosons turn into one, the amplitude for the process must include a factor of $g_{\rm YM}$. The key result we want to demonstrate is the following: there is a controllable $N \to \infty$ limit in which the physically relevant coupling is not $g_{\rm YM}$ but rather the 't Hooft coupling $\lambda = g_{\rm YM}^2 N$ that is kept finite.

To show this we consider a system of N coincident branes. The strength of the interactions between the open strings is controlled by the open string coupling constant g_o: each time an open string splits into two strings, or two strings join to form a single one, the amplitude for the process must include a factor of g_o. In the low energy limit we find the $SU(N)$ gauge theory and a decoupled $U(1)$ theory that we can ignore. Since the gauge bosons are simply massless open strings, the coupling constant $g_{\rm YM}$ of the resulting gauge theory coincides with the open string coupling g_o. In the argument that follows we use open strings to derive a result in the gauge theory. Accordingly, we will simply write $g_{\rm YM}$ for the open string coupling constant.

We examine the propagation of an open string whose ends lie on the i and j branes, with $i \neq j$. To find the quantum-mechanical amplitude for propagation of this string from some initial to some final conditions we must sum over all possible intermediate states that are consistent with beginning and ending with the $[ij]$ string. This sum, of course, is very complicated to do explicitly, but we are only going to determine the $g_{\rm YM}$ and N dependence of the various contributions. The result will apply to the amplitude of propagation of a gauge boson in the $SU(N)$ gauge theory.

The various contributions can be organized with diagrams where we show the evolution of the string. The simplest diagram is one where the string does nothing – it just propagates freely from the initial to the final condition. Figure 23.3(a) shows the strip produced by the motion of the open string, with the edges labeled by the branes i and j. This diagram has no interactions and no N-dependence. We associate to it an amplitude

$$A_0 = c_0 \,, \tag{23.37}$$

where c_0 is some constant independent of $g_{\rm YM}$ and N.

The next diagram includes interactions. The simplest possibility is that the string splits and rejoins, as shown in Figure 23.3(b). The two interactions, indicated by the heavy dots,

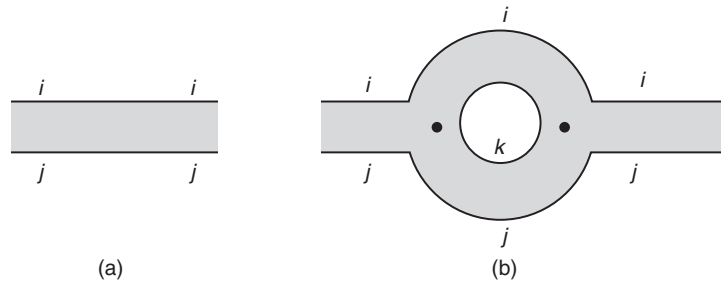

Fig. 23.3 **(a) Strip created by the free evolution of an [*ij*] string. (b) Diagram produced when the [*ij*] string splits and then rejoins at the *k*th brane.**

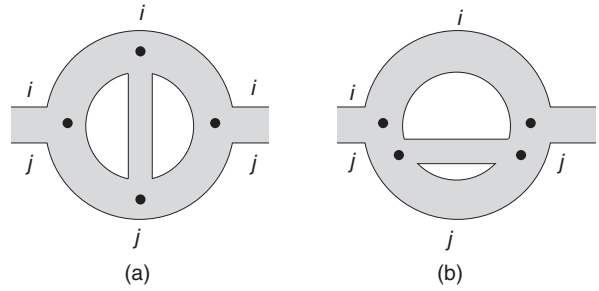

Fig. 23.4 **Two diagrams, each with four interaction points, obtained by adding a strip to diagram (b) in Figure. 23.3.**

give a factor $(g_{\rm YM})^2$. The inner boundary of the diagram has a label k, indicating that the string has split on the kth brane. Since that brane can be any of the N branes available and we must sum over all possibilities, the diagram also has a factor of N. Including a constant c_1, we write

$$A_1 = c_1 g_{\rm YM}^2 N\,. \tag{23.38}$$

A pattern emerges: each boundary that is a closed line contributes a factor of N, since it represents an open string endpoint that can be at any of the N branes. Suppose we add another strip to the diagram of Figure 23.3 (b). One simple way to do that will introduce two new interaction points and create one new boundary, as shown by the two diagrams of Figure 23.4. If this happens we get an extra factor of $g_{\rm YM}^2 N$ in the amplitude. We thus have

$$A_2 = c_2 (g_{\rm YM}^2 N)^2\,. \tag{23.39}$$

As long as we keep adding strips that create a new boundary we get amplitudes of the form

$$A_p = c_p (g_{\rm YM}^2 N)^p\,. \tag{23.40}$$

The amplitude A obtained by adding all of the above contributions is

$$A = \sum_{n=0}^{\infty} A_n = \sum_{n=0}^{\infty} c_n (g_{\rm YM}^2 N)^n\,. \tag{23.41}$$

We see here the emergence of the 't Hooft coupling constant λ defined by

$$\lambda \equiv g_{\rm YM}^2 N\,. \tag{23.42}$$

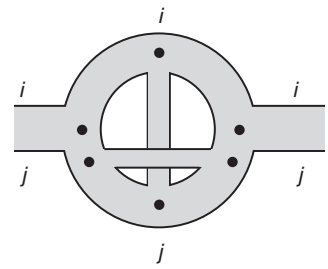

Fig. 23.5 A non-planar diagram obtained by adding a strip to any one of the two diagrams in Figure 23.4. Note that there is just one internal boundary.

The 't Hooft coupling controls the convergence of (23.41). We write

$$A = \sum_{n=0}^{\infty} c_n \lambda^n = f_0(\lambda)\,. \tag{23.43}$$

To find the complete amplitude we must add to A the contributions that arise when we add strips that do not create a new boundary. The simplest example involves adding a strip to any of the two diagrams of Figure 23.4 to obtain the diagram in Figure 23.5. We have introduced two new interactions, as usual, but the number of boundaries has been reduced by one. The two internal boundaries have become a single one, as you can check by moving the tip of a pencil along the boundary. This graph is actually non-planar: it cannot be drawn on the plane without making the strips cross. In fact a new strip will either increase the number of boundaries by one or decrease the number of boundaries by one. Each time we add a strip that decreases the number of boundaries we get an extra factor of $g_{\rm YM}^2/N$ or, equivalently, a factor of λ/N^2. Adding to A the contributions from diagrams that include all possible numbers of boundary-decreasing strips we find the full amplitude $\mathcal{A}$:

$$\mathcal{A} = f_0(\lambda) + f_2(\lambda)\cdot\frac{1}{N^2} + f_4(\lambda)\cdot\frac{1}{N^4} + \cdots\,. \tag{23.44}$$

When N is large and λ is fixed, the above is an expansion of the amplitude in powers of the small parameter $1/N^2$. The coefficient of N^{-2k}, with $k = 0, 1, 2, \ldots$, is controlled by the value of λ. If λ is small, each coefficient has an expansion in powers of λ. The limit $N \to \infty$ with $\lambda = g_{\rm YM}^2 N$ fixed requires $g_{\rm YM} \to 0$ in a controlled fashion. In this limit only the first term of the series contributes. The 't Hooft coupling λ fully controls the theory in the $N \to \infty$ limit.

23.5 Gravitational effects of massive sources

We examined a large-N gauge theory using open strings on N D-branes. We now ask: when are the gravitational effects of these D-branes important? After all, D-branes have energy (due to their tension) and must curve spacetime. More generally, we can ask: at

what distance scales are the gravitational effects of massive objects important? After we discuss this point, we focus on the case of branes wrapped around extra dimensions.

The relevant quantity to consider is the gravitational potential energy per unit mass, that is, the gravitational potential V created by a mass M at a distance r from the mass:

$$V \simeq -\frac{G^{(D)}M}{r^{D-3}} \,. \tag{23.45}$$

Here D is the number of spacetime dimensions. Since V is dimensionless in natural units, the numerator defines a characteristic length scale R:

$$R^{D-3} \equiv G^{(D)}M \,. \tag{23.46}$$

The scale R coincides, up to factors of order one, with the Schwarzschild radius of a mass M black hole in arbitrary dimension D. Indeed, our derivation of the radius of a four-dimensional black hole in Section 22.6 follows a similar logic. With R so defined, the potential (23.45) can be written as

$$V \simeq -\Big(\frac{R}{r}\Big)^{D-3} \,. \tag{23.47}$$

We see that gravitational effects are negligible for $r \gg R$. Gravitational effects of point masses are important at scales of order R. If R, however, is much smaller than the size of the mass M, gravitational effects may be completely negligible. A billion ton asteroid, for example, has $R \sim 10^{-15}$m. There are no significant gravitational effects at this scale nor at larger scales.

Consider now a number N of Dp branes wrapped around a p-dimensional compact space of volume V_p. The tension T_p of a Dp-brane is given by (18.57):

$$T_p \simeq \frac{1}{g}\frac{1}{(\sqrt{\alpha'})^{p+1}} \,. \tag{23.48}$$

The mass of the branes is therefore

$$M = NT_p \cdot V_p \simeq N \cdot \frac{V_p}{g(\sqrt{\alpha'})^{p+1}} \,. \tag{23.49}$$

The dimensionally reduced spacetime is $D-p$ dimensional. Moreover, in this spacetime the branes appear as a point source of mass M. Using (23.46) the characteristic size R of the system is

$$R^{D-p-3} = G^{(D-p)}M = \frac{G^{(D)}}{V_p}M \simeq G^{(D)}N \cdot \frac{1}{g(\sqrt{\alpha'})^{p+1}} \,. \tag{23.50}$$

Note the cancellation of the volume factor V_p. We recall that $G^{(D)} \sim g^2(\sqrt{\alpha'})^{D-2}$ (Section 13.4), so we get

$$R^{D-p-3} \simeq gN(\sqrt{\alpha'})^{(D-p-3)} \,. \tag{23.51}$$

We write this result as

$$\left(\frac{R}{\sqrt{\alpha'}}\right)^{D-p-3} \simeq gN \ . \tag{23.52}$$

As $gN \to 0$ we have $R \to 0$ and the gravitational effects of the N D-branes go to zero. In fact, as R becomes smaller than the string length $\sqrt{\alpha'}$ there is no scale at which the gravitational effects are relevant. As $g \to 0$ the mass (23.49) of the D-branes diverges like $1/g$, but the gravitational coupling G goes to zero like g^2. The overall effect, proportional to GM, vanishes. Note also the independence of the estimate (23.52) on the volume V_p of the compact dimensions. This suggests that the characteristic scale R is relevant even for configurations of infinite D-branes. In that case, however, one need not get a black hole, but rather some other geometry where R plays a significant role.

23.6 Motivating the AdS/CFT correspondence

In this section we motivate the AdS/CFT correspondence, a surprising equivalence between a supersymmetric $SU(N)$ gauge theory that arises at low energies on a set of N coincident D3-branes and a type IIB superstring theory in a background spacetime closely related to the gravitational background created by the D3-branes.

Although gauge theories naturally arise as low-energy limits of open string theory, the correspondence involves a *closed* string theory. Moreover, the closed strings do not live on Minkowski space, the space where the gauge theory lives. In fact, Minkowski space is, in some sense, the boundary of the space where the closed strings live. The correspondence is sometimes called a duality because the same physics is described by two different looking systems (the closed strings and the gauge theory). Since the $SU(N)$ gauge theory has maximal supersymmetry, its physics is different from that of QCD, a gauge theory with no supersymmetry. We are still lacking a string theory description of QCD, but the discovery of the AdS/CFT correspondence has given strong evidence that such a description exists.

It should be emphasized that the correspondence has not yet been proven. Rather, it was originally motivated by some heuristic arguments and has since been tested extensively. There are no grounds to suspect that it fails to hold. The heuristic arguments fall into two groups. They are either based on ideas of symmetry or they are based on a low energy limit.

Let us first briefly consider the symmetry argument. A conformal field theory in four-dimensional Minkowski space has a set of conformal symmetries that are generated by fifteen operators. Ten of these generators are the familiar Lorentz generators (six of them) and the spacetime translation generators (four of them). The other five include four that generate the so-called special conformal transformations and one that generates scale transformations. The fifteen operators define the four-dimensional conformal Lie algebra, which includes the Lorentz algebra (11.81) as a subalgebra. The conformal symmetries act on the field theory as a set of field transformations. These symmetries must also appear

on the string theory side of the correspondence. Indeed, the AdS_5 spacetime, which is part of the closed string background, has the property that the isometries of the space (the smooth, one-to-one maps of the space into itself that leave all distances invariant) are generated by fifteen operators that satisfy the same algebra as those that generate the conformal symmetry of the field theory. There is also a reason for the S^5 space: the supersymmetric Yang–Mills theory in question has a set of scalar fields and a set of fermions, the elements of which are rotated among each other by a set of symmetries that match the isometries of S^5. All in all, the $AdS_5 \times S^5$ spacetime carries as isometries the symmetries of the field theory, which provides some evidence for the correctness of the correspondence.

The argument based on a low energy limit is more direct. We will consider a set of N D3-branes, with N fixed and large, and we will dial up the string coupling constant g so that gN varies from very small $gN \ll 1$ to very large $gN \gg 1$. We will examine low energy limits in *both* extremes and extract some conclusions.

Let us start then with N coincident D3-branes at zero coupling g in flat ten-dimensional spacetime. We have free open strings on the branes and free type IIB closed strings on the spacetime. There are no interactions at all. Imagine now a minuscule increase of g to a fixed value such that $gN \ll 1$. Following our discussion in Section 23.5 we conclude that gravitational effects are negligible and we may continue to treat the D3-branes as if they were in flat space.

In a low energy limit we consider energies that are smaller than the string energy scale $1/\ell_s$:

$$E \ll \frac{1}{\sqrt{\alpha'}} \,. \tag{23.53}$$

Another way to think of the low energy limit is to imagine keeping all energies bounded while $\alpha' \to 0$:

$$E \leq E_0 \,, \quad \alpha' \to 0 \,. \tag{23.54}$$

For bounded energies, (23.53) holds for sufficiently small α'. In this limit, the massive states of the open strings on the D-branes are not accessible, so the physics on the branes is governed by the massless $U(N)$ Yang–Mills fields. As $\alpha' \to 0$, the closed string fields that propagate over the whole of the spacetime become free fields because the ten-dimensional Newton constant $G^{(10)} \sim g^2(\alpha')^4$ that governs their interactions (see (13.80)) goes to zero. Finally, the interactions between the spacetime fields and the $U(N)$ fields on the branes also go to zero because they too are controlled by $G^{(10)}$. The result is (1) a system of decoupled closed strings on the ten-dimensional Minkowski spacetime and (2) a supersymmetric four-dimensional $U(N)$ Yang–Mills theory. We learned in Section 15.5 that in a $U(N)$ Yang–Mills theory one gauge field actually decouples. The remaining, fully interacting gauge theory is $SU(N)$ Yang–Mills.

Consider now increasing the string coupling in the D3-brane system. As long as we consider low energy excitations, the excitations on the branes and the ones on the spacetime will continue to decouple. When $gN \gg 1$ gravitational effects are important. The branes carry energy and Ramond–Ramond charge. As a consequence, the N D3-branes are described by a nontrivial solution of the field equations for the massless fields of type IIB

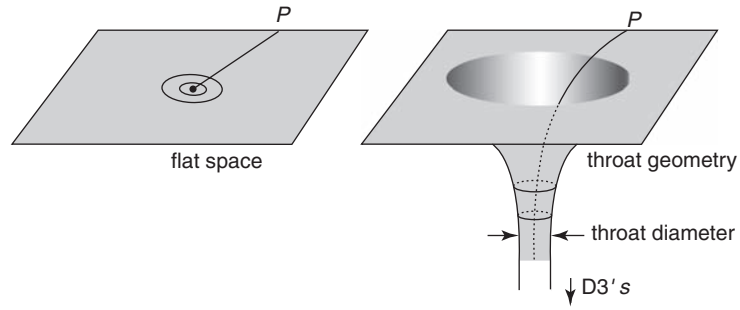

Fig. 23.6 **Left: the circumference of the circle surrounding the origin goes to zero as we approach the origin. Right: in the throat geometry, the origin has moved an infinite distance down the throat, and the circumferences of the circles which surround the origin approach a constant value. For the D3-brane solution, the surface represents the six-dimensional space transverse to the branes, and the circles on the throat represent five-spheres which surround the branes. The radial direction in the transverse space is represented by the line starting at *P* and going into the throat.**

string theory, much in the same way that a charged black hole is described. The solution which describes the D3-branes includes a horizon which lies at the end of an infinite throat.

To visualize such solution, consider first the origin of a flat world with two spatial dimensions (left side of Figure 23.6). The geometry of flat space is such that, at a distance r away from the origin, the circumference of the circle surrounding the origin is $2\pi r$. As we approach the origin, the circumference goes to zero. This changes dramatically if the geometry is that of a throat (right side of Figure 23.6). The origin is now an infinite distance down the throat. Moreover, down the throat, the circumference of a circle surrounding the origin approaches a constant called the circumference of the throat. Asymptotically, the throat becomes an infinite cylinder. The circle at the infinite "end" of the throat is called the horizon. The horizon is an infinite distance away from any point on the plane.

We can now return to our coincident D3-branes, which stretch along x^1, x^2, and x^3 but appear as a point along the six spatial coordinates $x^4, \ldots, x^9$ transverse to the brane. We need not focus on the longitudinal directions of the branes. In the transverse six-dimensional space the branes are surrounded by five-dimensional spheres. These are analogous to the circles that surround the origin in the two-dimensional example. A throat geometry with a horizon emerges in the transverse space. It takes an infinite distance to get to the horizon. The five-spheres that surround the horizon approach a constant volume as we travel down the throat. The radius R associated with this volume is called the radius of the horizon. It is worth noting that the geometry of the D3-brane solution is quite different from the geometry of the Schwarzschild black hole. In the latter, there is no infinite throat, and the horizon is a finite distance away from any point in the space.

The near-horizon geometry can be read directly from the metric which represents the gravitational solution: it turns out to be $AdS_5 \times S^5$. The D3-branes do not appear anymore in this geometry! The five-sphere was anticipated by our discussion of the throat geometry; its radius is the horizon radius R. The five-dimensional space AdS_5 arises from the four spacetime dimensions parallel to the branes plus the radial direction on the transverse

space. The five remaining transverse directions make up the sphere S^5. This rearrangement of the ten spacetime directions at the near horizon region is summarized as follows.

$$\begin{array}{cccccccccc} \overbrace{}^{\mathrm{AdS}_5} & & & & & \overbrace{}^{\mathrm{S}^5} & & & & \\ x^0 & x^1 & x^2 & x^3 & r & y^1 & y^2 & y^3 & y^4 & y^5 \\ \underbrace{x^0}_{\text{D3 tangential}} & \underbrace{x^1}_{\text{D3 tangential}} & \underbrace{x^2}_{\text{D3 tangential}} & \underbrace{x^3}_{\text{D3 tangential}} & x^4 & x^5 & x^6 & x^7 & x^8 & x^9 \\ & & & & \underbrace{}_{\text{D3 transverse}} & & & & & \end{array} \tag{23.55}$$

In here the six directions transverse to the brane became r – that went into the AdS space – and the five ys that make up the five-sphere.

A red-shift phenomenon that is familiar for black holes also occurs with the horizon. Finite energy excitations near the horizon are perceived as excitations of vanishingly small energy by an observer far away at infinity (the excitations are red shifted). As a result, low energy excitations for the observer at infinity can be of two types: *finite* energy excitations that emanate from the horizon or low energy (long wavelength) excitations far from the branes. There is evidence that these two types of excitations decouple. The ones that are far away are almost never captured by the horizon, which looks very small compared to the wavelength of the excitations. The excitations near the horizon cannot escape to infinity. As a result, the configuration is well approximated by two decoupled systems: (1) a system of low energy closed strings on flat space representing the far-away region and (2) a system of type IIB superstrings on the near-horizon $\mathrm{AdS}_5 \times S^5$ geometry.

We have described N D3-branes in two regimes, a first one for $gN \ll 1$ and a second one for $gN \gg 1$. The two regimes are connected by dialing up g, since we imagine N fixed and large. In the low energy limit each regime gave rise to two decoupled systems. In fact, in the low energy limit decoupling occurs for *all values* of gN and one of the decoupled systems is free closed strings. When $gN \ll 1$ the other decoupled system is $SU(N)$ Yang–Mills. Since this theory makes sense for all values of gN it is reasonable to expect that $SU(N)$ Yang–Mills is the other decoupled system for all values of the coupling gN. Alternatively, for $gN \gg 1$ the other decoupled system is IIB superstrings on $\mathrm{AdS}_5 \times S^5$. Since this theory makes sense for all values of gN it is reasonable to expect that IIB superstrings on $\mathrm{AdS}_5 \times S^5$ is the other decoupled system for all values of gN. See the diagramatic representation of the situation in Figure 23.7.

Consider first the decoupled system of closed strings in flat spacetime. As gN goes from very small to very big, the closed string system from the $gN \ll 1$ regime goes into the closed string system of the $gN \gg 1$ regime. If gN goes from very large to very small, the closed string system of the $gN \gg 1$ regime goes into the closed string system of the $gN \ll 1$ regime going through the *same* set of theories. In fact, in the low energy limit the closed strings are non-interacting for all finite g, so physically all these theories are the same.

Having matched one of the decoupled systems, we now look at the second one. As we mentioned before, we have two candidates for the second decoupled system: $SU(N)$ Yang–Mills and IIB superstrings on $\mathrm{AdS}_5 \times S^5$. The simplest possibility is that they are the same! Explicitly, as gN goes from very small to very big, the $SU(N)$ Yang–Mills theory

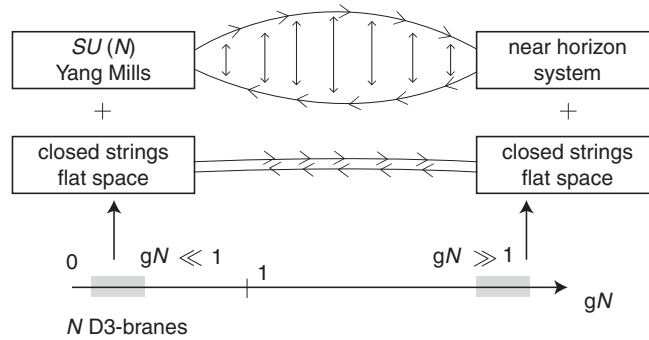

Fig. 23.7 **A diagramatic representation of the decoupling limits of a system of N D3-branes. To the left we have the branes with $gN \ll 1$, and its two decoupled systems above. To the right we have the branes with $gN \gg 1$, and its two decoupled systems above. If the two lines connecting the left and right top boxes represent equivalent systems, as indicated by the vertical identifications, one has the AdS/CFT correspondence.**

traces a line of theories that takes it to a complicated strong coupling regime that is, in fact, equivalent to the type IIB near horizon system that arises from the D3-branes at $gN \gg 1$. As gN goes from very large to very small, that near horizon system traces a line of theories that takes it to a complicated gravitational regime that is, in fact, equivalent to the weakly coupled $SU(N)$ Yang–Mills system that arises from the D3-branes at $gN \ll 1$. Moreover, the two lines of theories just discussed are in fact equivalent, as indicated by the vertical identifications in the figure. If this plausible scenario is realized, for any value of g and N there are two equivalent descriptions of the same physics, one using a gauge theory and another using the near horizon system, that is, closed superstrings on $AdS_5 \times S^5$. This is the AdS/CFT correspondence.

23.7 Parameters in the AdS/CFT correspondence

Let us discuss the parameters of the two theories in the AdS/CFT correspondence. In an $SU(N)$ Yang–Mills theory there are two dimensionless parameters: the coupling constant $g_{\rm YM}$ and the constant N. In IIB superstring theory on $AdS_5 \times S^5$ there are also two dimensionless parameters: the string coupling g and the radius $R/\sqrt{\alpha'}$ of the S^5 expressed in units of the string length. In summary:

$$\begin{aligned} &\text{Yang–Mills:} \quad g_{\rm YM}, \quad N, \\ &\text{IIB strings:} \quad g, \quad R/\sqrt{\alpha'}. \end{aligned} \tag{23.56}$$

Out of these four parameters, two of them are well defined in the original string theory system of D3-branes: the number N of D3-branes and the string coupling g. We need two relations to connect $g_{\rm YM}$ and R to g and N.

One relation is suggested from the limit $gN \ll 1$. We have stated earlier ((13.84)) that the open string coupling g_o is related to the closed string coupling g by $g_o^2 \sim g$. Since open

strings give rise to the Yang–Mills bosons of $SU(N)$, these bosons interact with $g_{\rm YM} \sim g_o$. It follows that $g^2_{\rm YM} \sim g_o^2 \sim g$. With a precise definition of the various couplings one has

$$g^2_{\rm YM} = 4\pi\, g\,. \tag{23.57}$$

The second relation is suggested from the region $gN \gg 1$. How is R, the horizon size, determined by the number of branes N and the string coupling? We studied a closely related question in Section 23.5, reaching the conclusion that the characteristic gravitational size of a system of D-branes is given by (23.52). In fact, this relation applies, and with $D - p - 3 = 10 - 3 - 3 = 4$ we get $R^4/\alpha'^2 \simeq g\,N$. The precise relation is

$$\frac{R^4}{\alpha'^2} = 4\pi\, g\, N\,. \tag{23.58}$$

The two equations above determine the relations between the gravitational and gauge parameters needed for the equivalence:

$$g = \frac{1}{4\pi}\, g^2_{\rm YM} \quad \text{and} \quad \frac{R^4}{\alpha'^2} = g^2_{\rm YM}\, N\,. \tag{23.59}$$

In terms of the 't Hooft coupling $\lambda = g^2_{\rm YM}\, N$, the equations above can be rewritten as

$$g = \frac{\lambda}{4\pi N} \quad \text{and} \quad \frac{R}{\sqrt{\alpha'}} = \lambda^{1/4}\,. \tag{23.60}$$

The string coupling is smaller than the 't Hooft coupling by a factor of N. Moreover, in units of string length, the radius of the S^5 only depends on 't Hooft coupling.

The first relation in (23.59) shows that weak Yang–Mills coupling implies weak string coupling. Since theories are generally simpler to deal with at weak coupling, it may seem that the correspondence should be easy to test: both theories could be examined at weak coupling and the results compared. This argument is wrong on two accounts. First, the 't Hooft coupling is the relevant gauge coupling for N large, so we need small λ for simple gauge theory computations. Second, the ability to calculate and to make quantitative statements in the string theory side requires both weak coupling *and* large $R/\sqrt{\alpha'}$. If the sphere S^5 is large, its curvature is small and the superstring theory can be accurately approximated by a supergravity theory in which calculations are simpler. Since $R/\sqrt{\alpha'} = \lambda^{1/4}$ we have a clash: tractable IIB requires a large left-hand side while tractable gauge theory requires a small right-hand side.

Quick calculation 23.2 Let $\lambda' = \alpha'^2/R^4$ denote the string expansion parameter that must be small to have a tractable IIB theory. How are λ and λ' related? Consider a series like (23.44) in the Yang–Mills side of the correspondence for λ fixed and N large. What is the expansion parameter in the IIB side of the correspondence?

It is useful to bring in the precise value of the gravitational constant $G^{(10)} \sim g^2\alpha'^4$ in the superstring side of the correspondence:

$$16\pi G^{(10)} = (2\pi)^7\, g^2\alpha'^4\,. \tag{23.61}$$

We can now obtain a couple of interesting results:

Quick calculation 23.3 Use the matching relations to show that

$$G^{(10)} = \frac{\pi^4 R^8}{2N^2} \,. \tag{23.62}$$

Since $G^{(10)} = \ell_{\rm P}^8$, where $\ell_{\rm P}$ is the ten-dimensional Planck length, $R/\ell_{\rm P} \sim N^{1/4}$. In Planckian units the radius of the sphere S^5 does not depend on the 't Hooft parameter.

Quick calculation 23.4 Show that the five-dimensional gravitational constant $G^{(5)}$ obtained after compactification on the S^5 of radius R is

$$G^{(5)} = \frac{\pi R^3}{2N^2} \,. \tag{23.63}$$

Despite the challenges, many nontrivial tests of the AdS/CFT correspondence have been carried out. The set of fields living on the gravitational side, for example, has been matched with the set of field operators that exist in the Yang–Mills theory. Supersymmetry is of help because it results in the existence of λ-independent quantities, known as protected observables. These can be calculated at zero coupling on the gauge theory side and compared with the gravitational predictions. Several such quantities have been compared successfully, and some new protected observables have been discovered. If we declare ourselves convinced that the AdS/CFT correspondence is correct, then the difficulty in testing it becomes a virtue: large-λ effects that are extremely difficult to compute in the gauge theory are calculable in the ten-dimensional supergravity theory!

Much of the recent work on the AdS/CFT correspondence has dealt with situations where the Yang–Mills theory has less, or no, supersymmetry. Such extensions are needed to obtain a correspondence that could apply to QCD. In addition, much work has also been devoted to develop other kinds of large-N limits, where the correspondence can be tested directly at weak coupling.

23.8 Hyperbolic spaces and conformal boundary

In this section we begin our preparation for the study of the geometry of anti-de Sitter spaces. We are already familiar with Minkowski space M_{n+1}, a flat space with one time dimension and n space dimensions. Anti-de Sitter space AdS_{n+1} also has one time dimension and n space dimensions, but it is curved. In fact, it is a negatively curved spacetime. De Sitter space dS_{n+1} has one time dimension and n space dimensions and is positively curved. Curved spacetimes are challenging to visualize, so we begin by discussing curved spaces that have no time direction. In such spaces the metric is positive definite and, consequently, all vectors have positive length squared (they are spacelike). We will focus on hyperbolic spaces, negatively curved spaces without a time direction. Hyperbolic spaces will help us understand anti-de Sitter spacetimes.

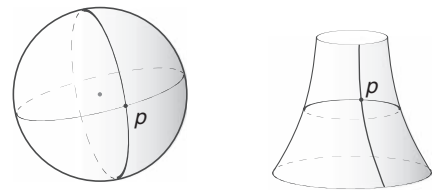

Fig. 23.8 **Left: a sphere, positively curved at p, with two lines that are shown to bend towards the same side of the surface – the interior. Right: a manifold negatively curved at p, as implied by the existence of two lines through p that bend to opposite sides of the surface.**

The unit sphere S^2 is the simplest example of a space of constant positive curvature. We can visualize it inside $\mathbb{R}^3$, with coordinates x^1, x^2, and x^3, as the surface $(x^1)^2 + (x^2)^2 + (x^3)^2 = 1$. This is an isometric embedding: the metric induced on the surface by the ambient metric (recall Section 6.2) is the familiar round metric on the sphere. It is generally easy to tell by inspection if a two-dimensional surface embedded in $\mathbb{R}^3$ is positively or negatively curved at a point p. Consider a vector $\vec{n}_p$ normal to the surface at p and a plane that contains $\vec{n}_p$. The intersection of the plane with the surface defines a line through p. By rotating the plane about the axis along $\vec{n}_p$ we get a family of lines that lie on the surface and go through p in all possible directions. A surface has positive curvature at p if all of the above lines bend towards the same side of the surface at p. A surface has negative curvature if one can find two lines that bend towards opposite sides of the surface at p. In Figure 23.8 we show a sphere and two lines through a point p. In fact, any line drawn in the manner explained above is a circle centered at the center of the sphere, and they all bend towards the interior of the sphere. We also show, to the right, a piece of a surface that is negatively curved at p. The horizontal circle bends in while the vertical line bends out.

The classic example of a two-dimensional space with constant negative curvature is hyperbolic space $\mathbb{H}_2$. It is actually a space with infinite volume and without a boundary. The first complication with this space is that, as opposed to the sphere S^2, it is not possible to fully embed it isometrically in $\mathbb{R}^3$. Certain portions of $\mathbb{H}_2$ can be embedded, and locally they look like the surface shown on the right side of Figure 23.8. Just like $\mathbb{R}^2$ and S^2, the space $\mathbb{H}_2$ is homogeneous: there are no special points, or equivalently, any point p can be moved to any point q by an isometry. For $\mathbb{R}^2$ the isometry is a translation and for S^2 the isometry is a rotation.

While $\mathbb{H}_2$ cannot be presented isometrically as a surface in $\mathbb{R}^3$ it can be presented isometrically as a surface in three-dimensional *Minkowski* space M_3. This presentation is very neat, but we lose perception of distances. Concretely, we have

$$\begin{aligned} \text{ambient metric:} \quad & ds^2 = -(dz)^2 + (dx^1)^2 + (dx^2)^2 \,, \\ \text{constraint:} \quad & -z^2 + (x^1)^2 + (x^2)^2 = -R^2 \,. \end{aligned} \tag{23.64}$$

The first equation gives the metric of M_3, showing that z is the time coordinate and that x^1 and x^2 are the space coordinates. The second equation is the constraint that defines the surface. Since we have one constraint on points in a three-dimensional space, the surface

is two-dimensional. The constraint indicates that $z^2 \geq R^2$, so the surface has two disconnected leafs: it exists for $z \geq R$ and for $z \leq -R$. The space $\mathbb{H}_2$ includes only one leaf; we can choose that to be the leaf $z \geq R$.

Since we claim that the surface has no time direction, no tangent to the surface can be timelike. This can be verified as follows. First note that the constraint equation implies that the position vector $v^\mu = (z, x^1, x^2)$ for a point on the surface is timelike. In fact, $v^\mu v_\mu = v \cdot v = -R^2$. A tangent is an infinitesimal vector δv^μ such that $v^\mu + \delta v^\mu$ still belongs to the surface: $(v + \delta v) \cdot (v + \delta v) = -R^2 + \mathcal{O}(\delta v^2)$. This means that $v \cdot \delta v = 0$. Since two timelike vectors cannot be orthogonal, the tangent vector δv must be spacelike. This is what we wanted to show.

Quick calculation 23.5 Why is the dot product of two timelike vectors always different from zero?

Quick calculation 23.6 Consider the point $(z, x^1, 0)$ on the hyperboloid. How are z and x^1 related? Exhibit two orthogonal (spacelike) tangents at the point. Note that this implies that all tangent vectors are spacelike.

It is worth noting that for surfaces defined by vectors v with constant length, all tangents δv at v satisfy $v \cdot \delta v = 0$. This means that the vector v is orthogonal to the surface. Moreover, since v^2 is nonvanishing, v cannot be a tangent at v. Since the normal to the surface is timelike and the ambient space has no two orthogonal timelike directions, the surface cannot have a timelike tangent vector.

Lorentz transformations preserve norm so acting on a vector that ends on the surface will give another vector that ends on the surface. In fact, any two vectors v_1 and v_2 that end on the surface can be mapped into each other by a Lorentz transformation: there are Lorentz transformations L_1 and L_2 that take both v_1 and v_2 to the vector $(R, \vec{0})$, so $L_1 v_1 = L_2 v_2$ and $v_2 = (L_2)^{-1} L_1 v_1$. Since Lorentz transformations preserve the ambient metric they act on the surface as isometries. All in all, this means that the surface is a homogeneous space, the surface has no special points.

Our next task is to determine the induced metric on the surface. Since this is not more complicated with higher number of dimensions, let us consider the case of $\mathbb{H}_n$, described as a surface in M_{n+1}. Using an index i that runs from 1 to n, we write

$$\begin{aligned} &\text{ambient metric:} \quad ds^2 = -(dz)^2 + dx^i dx^i\,, \qquad i = 1, 2, \ldots, n; \\ &\text{constraint:} \quad -z^2 + x^i x^i = -R^2\,. \end{aligned} \tag{23.65}$$

As usual, repeated indices imply summation. Again, one selects the leaf $z \geq R$. Points in the ambient space are denoted by coordinates $(z, x^1, x^2, \ldots, x^n) = (z, \vec{x})$. It is amusing to note that the mass-shell condition $p^2 = -(p^0)^2 + (p^1)^2 + (p^2)^2 + (p^3)^2 = -m^2$ defines the hyperbolic space $\mathbb{H}_3$ embedded in four-dimensional momentum space.

A nice form of the metric on $\mathbb{H}_n$ arises by a stereographic projection that takes the whole leaf $z \geq R$ to the *interior* of the ball B^n of radius R. The new coordinates are denoted by

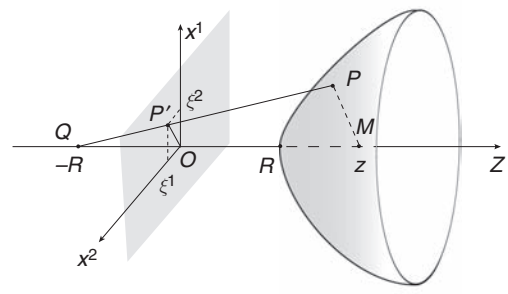

Fig. 23.9 The projection of a point P on the hyperboloid to a point P' on the screen $z = 0$. The ξ coordinates of P are equal to the x coordinates of P'.

ξ^i, with $i = 1, \ldots, n$. We use the vector notation $\vec{\xi} = (\xi^1, \xi^2, \ldots, \xi^n)$, and r is defined to be the radial coordinate in ξ variables:

$$r^2 \equiv \vec{\xi} \cdot \vec{\xi} = (\xi^1)^2 + (\xi^2)^2 + \cdots + (\xi^n)^2. \tag{23.66}$$

For a given point P on the hyperboloid consider the line from P to the point $Q = (-R, \vec{0})$ and the intersection P' of this line with the hyperplane $z = 0$, called the screen. We then set $\xi^i(P) = x^i(P')$, namely, the ξ coordinate of a point on the hyperboloid is declared equal to the x coordinate of its image on the screen. The three-dimensional representation in Figure 23.9 illustrates the construction for the case of $\mathbb{H}_2$. It follows from the projection that $\vec{\xi}$ is parallel to $\vec{x}$. Moreover, the ratio of their lengths is determined by the length r of $\vec{\xi}$: all points in a circle of radius r on the screen come from points with the same value of $|\vec{x}|$. We thus write

$$x^i = \rho(r)\, \xi^i \quad \rightarrow \quad |\vec{x}| = \rho\, r\,, \tag{23.67}$$

where $\rho(r)$ is a function to be determined. Using the similar triangles QOP' and QMP we have the relation

$$\frac{r}{R} = \frac{\rho\, r}{R + z} \quad \rightarrow \quad \frac{1}{R} = \frac{\rho}{R + \sqrt{R^2 + \rho^2 r^2}}\,, \tag{23.68}$$

where we have used

$$z^2 = R^2 + \vec{x} \cdot \vec{x} = R^2 + \rho^2 r^2\,. \tag{23.69}$$

Soving for ρ we find

$$\rho = \frac{2R^2}{R^2 - r^2} = \frac{2}{1 - \frac{r^2}{R^2}}\,. \tag{23.70}$$

Since ρ must be positive, the condition $r^2 < R^2$ guarantees that we get all possible values of ρ and no negative ones. Back in (23.67) we have that x^i and ξ^i coordinates are related by

$$x^i = \frac{2R^2}{R^2 - r^2}\, \xi^i\,. \tag{23.71}$$

To obtain the metric, we first calculate z as a function of r using (23.69) and (23.70). We find

$$z = R\,\frac{R^2 + r^2}{R^2 - r^2}\,. \tag{23.72}$$

Noting that $dr^2 = 2\xi^i d\xi^i$ a short computation gives

$$dz = \frac{4R^3}{(R^2 - r^2)^2}\,\xi^i d\xi^i\,. \tag{23.73}$$

Taking differentials of (23.71) we obtain:

$$dx^i = \frac{2R^2}{R^2 - r^2}\,d\xi^i + \frac{4R^2}{(R^2 - r^2)^2}\,\xi^i\xi^j d\xi^j\,. \tag{23.74}$$

Squaring dx^i and simplifying,

$$dx^i dx^i = \frac{4R^4}{(R^2 - r^2)^2}\,d\xi^i d\xi^i + \frac{16R^6}{(R^2 - r^2)^4}\,(\xi^i d\xi^i)^2\,. \tag{23.75}$$

Finally, we combine the results in (23.73) and (23.75) into $ds^2 = -dz^2 + dx^i dx^i$, and immediately find

$$ds^2 = \frac{4\,d\xi^i d\xi^i}{\left(1 - \frac{r^2}{R^2}\right)^2}\ , \tag{23.76}$$

where we recall that $r = \sqrt{\xi^i\xi^i} < R$. It is sometimes convenient to scale the disk to unit radius. For this we simply let $\xi^i \to R\xi^i$, which also gives $r \to R\,r$. The metric then becomes the standard

$$\mathbb{H}_n \text{ metric}: \ ds^2 = \frac{4R^2\,d\xi^i d\xi^i}{(1 - r^2)^2},\ i = 1, \ldots, n, \quad r = \sqrt{\xi^i\xi^i} < 1\,. \tag{23.77}$$

In this presentation the ξ coordinates have no units, the length scale is provided by the explicit factors of R, the radius of curvature of the space. Hyperbolic space has infinite volume. In fact a line from $r = 0$ to $r = 1$ has infinite length. To check this consider a line that goes from $\xi^1 = 0$ to $\xi^1 = \bar{r}$, with all other $\xi^i = 0$. Then the length $\ell(\bar{r})$ of this curve is

$$\ell(\bar{r}) = \int_0^{\bar{r}} \frac{2R\,d\xi^1}{1 - (\xi^1)^2} = R\,\ln\!\left(\frac{1 + \bar{r}}{1 - \bar{r}}\right). \tag{23.78}$$

Indeed, as $\bar{r} \to 1$ the length $\ell(\bar{r})$ diverges.

It seems difficult to speak of the boundary $r \to 1$ of $\mathbb{H}_n$ since it lies an infinite distance away from any point. There is, however, the notion of a *conformal boundary* that can be made precise and carries very interesting information. Given a metric ds^2 we modify it by multiplication by an extra factor $\Omega^2 > 0$ to obtain a new metric $ds'^2 = \Omega^2 ds^2$. The factor Ω^2 is chosen so that in the new metric the distance from any point to all boundary points is

finite. The conformal boundary is the boundary in the new metric ds'^2. This boundary can be examined clearly since it lies at a finite distance. The conformal boundary is only defined up to additional multiplication by factors Ω'^2 that keep all boundary points a finite distance away. While statements about the precise length or volume of the conformal boundary are meaningless, there is some invariant information. For some spaces the conformal boundary is always a point. For some spaces the conformal boundary can have an extent. If we are dealing with *spacetimes* with nontrivial conformal boundaries the character of tangent vectors to the boundary (spacelike, timelike, or null) is invariant information since the sign of the norm of a vector cannot be changed by multiplication by $\Omega^2 > 0$.

Let us consider, as a first example, two-dimensional flat space $\mathbb{R}^2$ with coordinates $-\infty \le x, y \le \infty$ and flat metric $ds^2 = (dx)^2 + (dy)^2 = (dr)^2 + r^2(d\theta)^2$, where r and θ are the standard polar coordinates. We produce a conformally related metric by multiplication by the square of $\mathcal{C}(r,\theta) > 0$:

$$ds'^2 = \mathcal{C}^2(r,\theta)ds^2 = \mathcal{C}^2(r,\theta)\big((dr)^2 + r^2(d\theta)^2\big)\,. \tag{23.79}$$

If lines from the origin to infinity along constant θ have finite length on the new metric we must have

$$\int_0^\infty \mathcal{C}(r,\theta)\,dr < \infty\,. \tag{23.80}$$

Integrating over θ and exchanging the order of integration,

$$\int_0^\infty \frac{dr}{r}\int_0^{2\pi} \mathcal{C}(r,\theta)\,r\,d\theta < \infty\,. \tag{23.81}$$

Noting that the integral over θ gives precisely the length $\ell(r)$ of the constant r circle, we write

$$\int_0^\infty \frac{dr}{r}\,\ell(r) < \infty\,. \tag{23.82}$$

If $\ell(r)$ has a nonzero limit as $r \to \infty$ the integral would diverge. So we have learned that $\ell(r) \to 0$ as $r \to \infty$ in any suitable conformal metric. The length of circles approaching infinity goes to zero, and therefore the conformal boundary is just a point. The conformally related space is a sphere with a missing point. In Section 22.6 we will examine the construction of the "Riemann sphere" along related lines. One can also prove that the conformal boundary of $\mathbb{R}^n$ with $n > 2$ is always a point.

We can now turn to hyperbolic space (23.77). It is clear that we can get something quite simple by multiplication by the conformal factor $(1 - r^2)^2/(4R^2)$:

$$ds'^2 = \frac{1}{4R^2}(1 - r^2)^2\,ds^2 = d\xi^i d\xi^i\,. \tag{23.83}$$

The conformally related space is just the interior $\xi^i\xi^i < 1$ of the unit ball B^n with the flat constant metric. The conformal boundary is the set $\xi^i\xi^i = 1$, that is, the unit sphere S^{n-1}. It is interesting that the conformal boundary of hyperbolic space is much more substantial than that of flat space.

23.9 Geometry of AdS and holography

In this section we examine in some detail the geometry of AdS spaces in order to appreciate the holographic aspects of the AdS/CFT correspondence. In optics, a hologram is a two-dimensional plate onto which an image of a three-dimensional object has been recorded. In photography, one directly focuses the image of an object into the film. In holography, one records an interference pattern between coherent light reflected from the object and coherent light from a reference beam. This recorded interference pattern is still two-dimensional, but it contains much more information than a photograph. It contains information about all three dimensions of the object. The pattern allows you to view a three-dimensional image that exhibits parallax: the image looks different depending on the direction of the observer.

In gravitational physics a system that extends over a macroscopic region of space is said to be holographic if all of its physics can be represented by a theory that lives on the boundary of this region. Moreover, one requires that the boundary theory should not contain more than one degree of freedom (with a finite number of states) per Planck area. Holography is motivated by the physics of black holes: the entropy of a black hole is not proportional to the volume enclosed by its horizon but rather to the area of the horizon. Moreover, as discussed previously (equation (22.124)), this entropy is reproduced by assuming that the horizon carries a degree of freedom per horizon element of Planck area. The AdS/CFT correspondence is a more concrete realization of holography. There is a precise sense in which the four-dimensional space where the Yang–Mills theory lives can be viewed as a boundary of AdS_5. The $SU(N)$ theory, which captures all the physics of the interior of the ten-dimensional spacetime, provides a holographic description of the gravitational world. Moreover, one can roughly argue that the holographic bound on the degrees of freedom at the boundary holds.

The space AdS_{n+1} is usually defined as a surface embedded in a flat space $\mathbb{R}^{2,n}$ with two time coordinates u and v and n space coordinates x^i:

$$\begin{aligned} &\text{ambient metric:} \quad ds^2 = -(du)^2 - (dv)^2 + dx^i dx^i\,, \quad i = 1, \ldots, n; \\ &\text{constraint:} \quad -u^2 - v^2 + x^i x^i = -R^2\,. \end{aligned} \tag{23.84}$$

On the space $\mathbb{R}^{2,n}$ there are generalized Lorentz transformations, linear transformations that preserve the metric. The constraint states that a vector $V = (u, v, \vec{x})$ belongs to the surface if $V \cdot V = -R^2$, where the dot product uses the ambient metric. The vector V is timelike. The above equations are analogous to those that define hyperbolic space in (23.8), except that now we have two time directions. Since V is defined by a length condition, it is normal to the surface. Given that V is timelike and the ambient space has two orthogonal timelike directions, the surface must contain one timelike direction. In fact, this direction can be readily visualized. Fix a specific point $(u_0, v_0, \vec{x}_0)$ on the surface. Clearly,

$$u_0^2 + v_0^2 = R^2 + \vec{x}_0 \cdot \vec{x}_0\,. \tag{23.85}$$

Consider now the circle defined by all values of u and v that satisfy

$$u^2 + v^2 = R^2 + \vec{x}_0 \cdot \vec{x}_0\,. \tag{23.86}$$

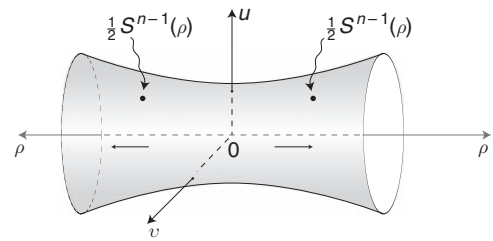

Fig. 23.10 **AdS$_{n+1}$ space described as a two-dimensional surface in which each point represents *half* of the $S^{n-1}(\rho)$ sphere. The horizontal axis has ρ increasing from zero to infinity *both* to the left and to the right. The half spheres at points with equal values of u and v (shown in figure) must be glued to form a single $S^{n-1}(\rho)$ sphere.**

This circle lies on the surface and goes through the point $(u_0, v_0, \vec{x}_0)$. Moreover, the tangent vectors to the circle are timelike everywhere since they are vectors with components only along u and v.

Anti-de Sitter space is a homogeneous space: no point is special and any point can be mapped into any other point by a transformation that is an isometry. Indeed, two vectors V_1 and V_2 that have the same norm $V_1 \cdot V_1 = V_2 \cdot V_2 = -R^2$ can always be mapped into one another by a generalized Lorentz transformation (you may enjoy proving this statement). In order to convince yourself that the surface does not contain two independent timelike directions anywhere it suffices to verify that two orthogonal timelike directions do not exist at one point. This is readily done.

Quick calculation 23.7 Consider the point $(R, 0, \vec{0})$ on the surface. Show that the surface does not contain two orthogonal timelike directions at this point.

We can visualize AdS space using a two-dimensional surface on which each point represents a sphere. We write the constraint equation as

$$u^2 + v^2 = R^2 + \vec{x} \cdot \vec{x} \tag{23.87}$$

and plot the space using axes for u, v, and $\rho = \sqrt{\vec{x} \cdot \vec{x}}$. A point on the two-dimensional surface shown in Figure 23.10 is determined by u and v, since then the value of ρ is fixed. To see the whole AdS space we must include at each point on the surface the sphere S^{n-1} defined by the points $\vec{x}$ that satisfy: $\vec{x} \cdot \vec{x} = \rho^2 = u^2 + v^2 - R^2$. As shown in the figure, the surface extends both to the left and to the right, and in *both* regions ρ varies from 0 to ∞! Had we plotted the space as the region that only goes to the right, you may reach the incorrect conclusion that the space has a boundary at $\rho = 0$, or equivalently on the circle $u^2 + v^2 = R^2$. This conclusion is suspect because on this circle the S^{n-1} spheres have zero radius. Since we represent the range of ρ twice, there are two points associated with each value of the pair u, v. Consistency requires that on top of each of these two points we place *half* of the S^{n-1} sphere. We cut the sphere in two and place half of it on top of each point, with the understanding that the halves are to be glued.

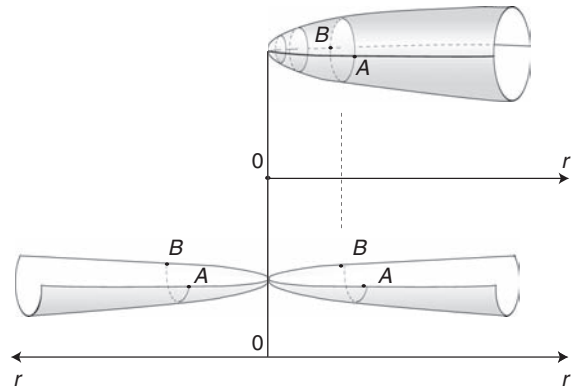

Fig. 23.11 **Top: a cigar as a collection of circles over the half-line $r \geq 0$, with the radii of the circles going to zero as $r \to 0$. Bottom: the same space represented with an r that grows from zero to infinity both to the right and to the left. Above each point there is now half a circle. The two half circles at r are glued to form a full circle.**

A lower-dimensional analog helps visualize the situation. Imagine a cigar represented as a circle S^1 over a semi-infinite line $r \geq 0$, as shown in the top part of Figure 23.11. The radius of the circle vanishes as $r \to 0$, where the cigar ends. The cigar, however, has no boundary at this point. A representation similar to the one we use for AdS, shown at the bottom, represents r in both directions and places half of each circle S^1 at symmetric points.

We identified a timelike direction on the AdS space, but in fact, we found something strange: closed timelike curves. Indeed, the (u, v) circles in (23.86) have timelike tangents everywhere and are closed. This is not the kind of space we want; in a universe with closed timelike curves one could travel for a while and get back before one's departure. What is sometimes called the "complete" AdS space, but we will simply call AdS space, is a space in which we unwrap the timelike circles in the space into open lines. The full AdS space can be imagined as a space rolled over the surface in Figure 23.10 an infinite number of times.

We now determine the metric on AdS_{n+1}. We begin by making the time coordinate explicit through the relations

$$u = z\,\cos t\,, \qquad v = z\,\sin t\,, \tag{23.88}$$

where we trade u and v for the time t and an additional coordinate z. It follows from these relations that

$$u^2 + v^2 = z^2\,, \quad \text{and} \quad (du)^2 + (dv)^2 = (dz)^2 + z^2(dt)^2\,. \tag{23.89}$$

We can now use these equations in (23.84) to find that the ambient metric and the constraint become

$$\begin{aligned} \text{ambient metric:} \quad & ds^2 = -z^2(dt)^2 - (dz)^2 + dx^i dx^i\,, \quad i = 1, \ldots, n\,; \\ \text{constraint:} \quad & -z^2 + x^i x^i = -R^2\,. \end{aligned} \tag{23.90}$$

Apart from the extra $-z^2(dt)^2$ appearing in the metric, the above are exactly the equations (23.8) for hyperbolic space. At any fixed time ($dt = 0$) the *spatial* geometry of the AdS space is that of $\mathbb{H}_n$. This is the way hyperbolic space appears in the anti-de Sitter geometry.

A little work remains now to find the full metric. As before we use ξ^i coordinates and write both x^i and z in terms of them. The value of z was determined in (23.72) and the portion $-(dz)^2 + dx^i dx^i$ of the metric was calculated in (23.76). We thus find that the metric in (23.9) becomes

$$ds^2 = -\frac{(R^2+r^2)^2}{(R^2-r^2)^2}\, R^2 (dt)^2 + \frac{4R^4\, d\xi^i d\xi^i}{(R^2-r^2)^2}\,. \tag{23.91}$$

As a last step we let $\xi^i \to R\xi^i$ and obtain the metric on AdS_{n+1}:

$$\mathrm{AdS}_{n+1}\ \text{metric}:\ ds^2 = R^2\left[-\Big(\frac{1+r^2}{1-r^2}\Big)^2 (dt)^2 + \frac{4\, d\xi^i d\xi^i}{(1-r^2)^2}\right]. \tag{23.92}$$

Here, as before, $r^2 = \xi^i \xi^i$ and $i = 1, 2, \ldots, n$. We already know from our analysis of $\mathbb{H}_n$ that at any fixed time, the distance from any point to the boundary $r \to 1$ is infinite. What is quite interesting is that it does not take infinite time for a light ray to reach the boundary. Assuming the light-ray travels in the ξ^1 direction, the condition $ds^2 = 0$ gives

$$\frac{1+r^2}{1-r^2}\, dt = \frac{2d\xi^1}{1-r^2} \;\to\; \frac{d\xi^1}{1+(\xi^1)^2} = \frac{1}{2}\, dt\,. \tag{23.93}$$

Integrating with $\xi^1 = 0$ for $t = 0$, we find $\xi^1 = \tan(t/2)$, which shows that the boundary point $\xi^1 = 1$ is reached at $t = \pi/2$.

To understand the conformal properties of the space and its boundary we use the metric (23.92) to define the conformally related

$$ds'^2 = \frac{(1-r^2)^2}{4R^2}\, ds^2 = -\Big(\frac{1+r^2}{2}\Big)^2 (dt)^2 + d\xi^i d\xi^i\,. \tag{23.94}$$

The metric ds'^2 describes a spatial n-dimensional ball (the interior of $\xi^i \xi^i = 1$) and a time coordinate. Near $r = 1$ we have

$$ds'^2 \simeq -(dt)^2 + d\xi^i d\xi^i\,, \quad \text{as} \quad r \to 1\,, \tag{23.95}$$

and the conformal boundary takes the form $\mathbb{R} \times S^{n-1}$, the $\mathbb{R}$ factor for the time and the S^{n-1} factor for the boundary $\xi^i \xi^i = 1$ of the hyperbolic spatial sections. For AdS_3 the boundary $\mathbb{R} \times S^1$ is the surface of a cylinder, and the full spacetime is inside of it (see Figure 23.12). It is a nontrivial fact that AdS spacetimes have a conformal boundary with both time and space directions. The conformal boundary of Minkowski space, for example, only contains null directions (see Problem 23.4).

For the AdS_5 spacetime the boundary is $\mathbb{R} \times S^3$. The gauge theory dual to the type IIB superstring background lives on this boundary. The scale ambiguity of the conformal

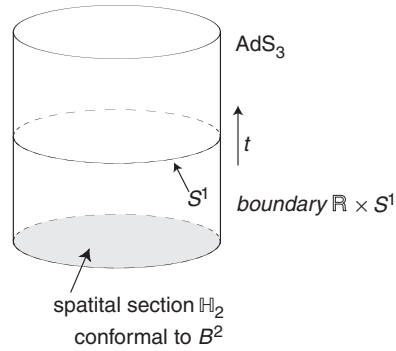

Fig. 23.12 **The conformal view of AdS$_3$. The conformal boundary is the surface $\mathbb{R} \times S^1$ of the cylinder, where $\mathbb{R}$ is time and S^1 is the conformal boundary of the spatial sections $\mathbb{H}_2$. The full AdS$_3$ space is conformal to the interior of the cylinder.**

boundary is consistent with the scale invariance of the field theory. The physics of the gauge theory does not care about the radius of the three-sphere. If we take the radius to infinity, the boundary is flat four-dimensional Minkowki spacetime.

We can obtain insight into AdS$_5$ holography by focusing on its fixed-time spatial geometry, which is that of $\mathbb{H}_4$. We calculate the volume and the boundary area of spherical subsets $\mathcal{S}(\bar{r})$ of $\mathbb{H}_4$ that include all points with $r \leq \bar{r}$. In the limit $\bar{r} \to 1$ $\mathcal{S}(\bar{r})$ approaches the full spatial section of AdS$_5$. The area of the boundary of $\mathcal{S}(\bar{r})$ is the "area" $A_3(\bar{r})$ of the three-sphere located at $r = \bar{r}$. The metric (23.92) tells us that physical length is obtained by multiplying coordinate length by $2R/(1 - r^2)$. It follows that

$$A_3(\bar{r}) = \Big[\frac{2R}{1-\bar{r}^2}\Big]^3 \mathrm{Vol}(S^3(\bar{r})) = \Big[\frac{2\bar{r}\,R}{1-\bar{r}^2}\Big]^3 \cdot 2\pi^2\,. \tag{23.96}$$

The volume $V_4(\bar{r})$ of $\mathcal{S}(\bar{r})$ is given by

$$V_4(\bar{r}) = \int_0^{\bar{r}} \Big[\frac{2R}{1-r^2}\Big]^4 \mathrm{Vol}(S^3(r))dr = 2\pi^2\,(2R)^4 \int_0^{\bar{r}} \frac{r^3 dr}{(1-r^2)^4}\,. \tag{23.97}$$

The integral is readily evaluated and gives

$$V_4(\bar{r}) = 2\pi^2\,(2R)^4 \frac{3\bar{r}^2-1}{12(1-\bar{r}^2)^3} = \frac{R}{6}\frac{3\bar{r}^2-1}{\bar{r}^3}\Big[\frac{2\bar{r}\,R}{1-\bar{r}^2}\Big]^3 \cdot 2\pi^2\,. \tag{23.98}$$

As expected, both the area $A_3(\bar{r})$ and the volume $V_4(\bar{r})$ diverge as $\bar{r} \to 1$. Their ratio, however, does not:

$$\frac{A_3(\bar{r})}{V_4(\bar{r})} = \frac{6}{R}\frac{\bar{r}^3}{3\bar{r}^2-1} \quad \to \quad \lim_{\bar{r}\to 1}\frac{A_3(\bar{r})}{V_4(\bar{r})} = \frac{3}{R}\,. \tag{23.99}$$

As the subsets grow without bound to cover $\mathbb{H}_4$, the boundary area and the bulk volume become proportional to one another! The constant of proportionality is the radius of curvature R of the AdS space. In flat space, a four-dimensional region of characteristic size L will have an area $A_3 \sim L^3$ and a volume $V_4 \sim L^4$. The ratio $A_3/V_4 \sim 1/L$ goes to zero as

the size of the region grows without bound. This is why holography is hard to realize, the boundary is too small to capture the physics of the bulk. Less so in AdS space, where the area to volume ratio does not approach zero.

We can elaborate further if we introduce a regulator that makes both boundary areas and number of degrees of freedom finite. Consider a small number $\delta \ll 1$ and the sphere $\bar{r} = 1 - \delta$. It follows from (23.96) that the area A_3^δ of this sphere is

$$A_3^\delta \equiv A_3(1-\delta) \simeq 2\pi^2 \frac{R^3}{\delta^3} \,. \tag{23.100}$$

Now let us consider the boundary theory at $\bar{r} = 1$. To estimate the number of degrees of freedom $N_{\rm dof}$ on this theory we take δ to define a short distance cutoff. We imagine having one degree of freedom per little cube of *coordinate* volume δ^3. Since the boundary has coordinate volume of order one, we have $1/\delta^3$ degrees of freedom. The boundary theory is made of $SU(N)$ fields, so we correct the estimate by multiplying by N^2; our final result being N^2/δ^3 degrees of freedom. We now verify that this number $N_{\rm dof}$ of degrees of freedom matches the holographic expectation of a degree of freedom per piece of surface of Planck size. We first note that

$$N_{\rm dof} \sim \frac{N^2}{\delta^3} \sim \frac{N^2 A_3^\delta}{R^3} \,. \tag{23.101}$$

Recalling (23.63), which tells us that $R^3 \sim G^{(5)} N^2$, we find

$$N_{\rm dof} \sim \frac{A_3^\delta}{G^{(5)}} \,. \tag{23.102}$$

This *is* the holographic expectation. If we take the total volume V of the space in the type II theory to be given by $V_4^\delta = V_4(1-\delta)$ times the volume $V_5 \sim R^5$ of the S^5, the number of degrees of freedom per unit volume behaves as

$$\frac{N_{\rm dof}}{V} \sim \frac{A_3^\delta}{V_4^\delta} \frac{1}{R^5\, G^{(5)}} \sim \frac{1}{R} \frac{1}{G^{(10)}} \,. \tag{23.103}$$

For large R the number of degrees of freedom per unit volume in the bulk is very small. If we increase the number of branes, R increases and we can make the ratio $N_{\rm dof}/V$ arbitrarily small. This bulk theory, with such small density of degrees of freedom, is string theory with a fixed string length $\sqrt{\alpha'}$.

23.10 AdS/CFT at finite temperature

The AdS_5 space supports black holes in the same sense that Minkowski space supports black holes: the geometry of the space is deformed by the presence of the hole but as we move away from the hole the spacetime metric approaches the original metric. A Schwarzschild black hole inside AdS_5 is described by the metric

$$ds^2 = -\Big(1 + \frac{r^2}{R^2} - \frac{r_0^2}{r^2}\Big)(dt)^2 + \Big(1 + \frac{r^2}{R^2} - \frac{r_0^2}{r^2}\Big)^{-1}(dr)^2 + r^2 d\Omega_3^2 \,. \tag{23.104}$$

The above metric describes a space that is *asymptotically* AdS. That is, as $r \gg r_0$

$$r \gg r_0 : \quad ds^2 = -\Big(1 + \frac{r^2}{R^2}\Big)(dt)^2 + \Big(1 + \frac{r^2}{R^2}\Big)^{-1}(dr)^2 + r^2 d\Omega_3^2 \,, \tag{23.105}$$

which is the metric of an AdS_5 space of radius R, written in a form different to the one we considered before (see Problem 23.5). Thus, in (23.104), R denotes the radius of the asymptotic AdS_5. The length parameter r_0 tells us that we have a black hole. One can write r_0 in terms of the mass M of the hole and the five-dimensional Newton's constant: $r_0^2 \sim G^{(5)} M$. In fact, as $R \to \infty$ the metric (23.104) becomes that of a black hole in five-dimensional Minkowski space with Schwarzschild radius r_0. Finally, $d\Omega_3^2$ denotes the metric on a three-sphere S^3 of unit radius.

The Schwarzschild radius r_+ of the AdS black hole is the value of r that makes the coefficient of $(dt)^2$ in the metric equal to zero:

$$1 + \frac{r_+^2}{R^2} - \frac{r_0^2}{r_+^2} = 0 \,. \tag{23.106}$$

Solving for r_+^2 we find

$$r_+^2 = \frac{R^2}{2}\Big(\sqrt{1 + \frac{4r_0^2}{R^2}} - 1\Big) \,. \tag{23.107}$$

For a fixed AdS scale R, the Schwarzschild radius r_+ is a function of r_0 or, equivalently, a function of the black hole mass. For $r_0 \ll R$ one can see that $r_+ \sim r_0$ and for $r_0 \gg R$ one finds $r_+ \sim \sqrt{r_0 R}$. In fact, r_+ is always smaller than r_0, as one can see by rewriting (23.107) in the form

$$\frac{r_+^2}{r_0^2} = \frac{2}{1 + \sqrt{1 + \frac{4r_0^2}{R^2}}} \leq 1 \,. \tag{23.108}$$

The scale R of the asymptotic AdS space is not a limit for the size of the black hole. A large black hole is one for which $r_+ \gg R$. A small black hole is one for which $r_+ \ll R$. A plot of r_+ as a function of r_0 is shown in Figure 23.13.

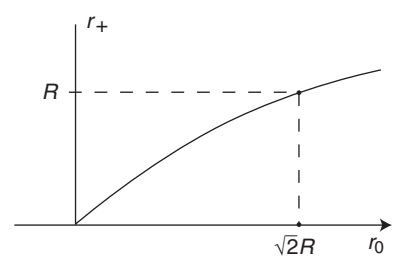

Fig. 23.13 The dependence of the Schwarzschild radius r_+ of the AdS black hole on the mass of the hole, encoded by r_0.

The black hole solution presented here has a Hawking temperature $\bar{T}_{\rm H}$. A calculation of this temperature gives:

$$\bar{T}_{\rm H} = \frac{R^2 + 2r_+^2}{2\pi r_+ R^2} \,. \tag{23.109}$$

While a justification of this result requires tools beyond those assumed in this text, it is worthwhile to know that this temperature emerges from a general formula as follows.

Quick calculation 23.8 Consider a metric $ds^2 = -f(r)(dt)^2 + (f(r))^{-1}(dr)^2 + \cdots$ with $f(r)$ a function of a radial coordinate that defines a horizon radius r_+ via $f(r_+) = 0$. The Hawking temperature $\bar{T}_{\rm H}$ associated with this gravitational solution is given by $\bar{T}_{\rm H} = f'(r_+)/(4\pi)$. Apply this result to (23.104) and derive (23.109).

The temperature (23.109) has interesting limits. For small black holes we find

$$\bar{T}_{\rm H} \simeq \frac{1}{2\pi r_+} \,, \quad \text{for} \quad r_+ \ll R \,. \tag{23.110}$$

In Minkowski space the temperature of a black hole is inversely proportional to the Schwarzschild radius (see (22.120)); small black holes are hot. We have recovered this result in (23.110) because for small black holes the curvature of the AdS is a negligible effect. More surprising is the behavior for large black holes. We then have:

$$\bar{T}_{\rm H} \simeq \frac{r_+}{\pi R^2} \,, \quad \text{for} \quad r_+ \gg R \,. \tag{23.111}$$

This is unusual: once the black hole is large enough, its temperature grows with its size. This is an important qualitative feature of black holes in AdS space. A sketch of the temperature $\bar{T}_{\rm H}$ as a function of r_+/R is shown in Figure 23.14.

Quick calculation 23.9 Show that for fixed radius of curvature R, all black holes have temperature $\bar{T}_{\rm H}$ that satisfies

$$\bar{T}_{\rm H} \geq T_0 = \frac{\sqrt{2}}{\pi} \frac{1}{R} \,, \tag{23.112}$$

where the lower bound T_0 is realized for $r_+/R = 1/\sqrt{2}$. This information is displayed in Figure 23.14.

It is reasonable to expect that the AdS/CFT correspondence extends to the case of a black hole inside the now asymptotic AdS space. Indeed, there is an obvious candidate for the dual gauge theory: $SU(N)$ Yang–Mills theory at finite temperature $\bar{T}_{\rm H}$, the temperature of the black hole. The metric (23.104) for fixed large r becomes

$$ds^2 \simeq \frac{r^2}{R^2}\Big[-(dt)^2 + R^2 d\Omega_3^2\Big] \,. \tag{23.113}$$

As in our discussion below (23.95) we see a boundary $\mathbb{R} \times S^3$, where the sphere is of radius R. While this radius is immaterial in the zero-temperature case, it is not immaterial now. The radius R is the radius of the sphere where the field theory at temperature $\bar{T}_{\rm H}$ lives.

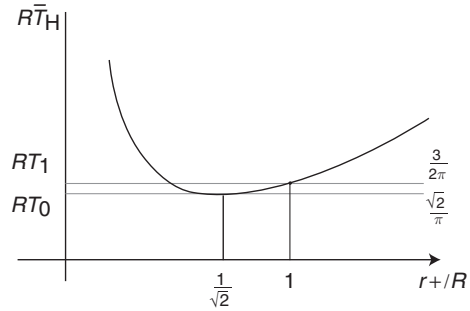

Fig. 23.14 **The Hawking temperature $\bar{T}_H$ of the AdS black hole (multiplied by the AdS radius R) sketched as a function of the Schwarzschild radius r_+ scaled down by R. For each value of the Hawking temperature above T_0 there are two black holes. The temperature T_1 defines the Hawking–Page transition.**

One cannot rescale with impunity the time coordinate t in (23.113) because at non-zero temperature the time coordinate carries information about temperature, which is an energy scale. The conformal ambiguity of the above metric still remains: if we multiply the factor in brackets by λ^2 this has the effect of changing the radius to λR and, by scaling time $t \to \lambda t$, changing the temperature to $\bar{T}_H/\lambda$. In the identification of the dual field theory the product $R\bar{T}_H$ of the radius and the temperature is the *only* invariant information.

As a test of the correspondence we calculate the entropy of the finite-temperature field theory and compare it with the entropy of the (large) black hole, hoping for an equality. Both of these numbers are finite (not infinite!). The field theory has a finite entropy density and lives on a space of finite volume, and the black hole has a horizon of finite area. Let us begin with the entropy S_{YM} of the field theory. We first recall that the entropy *density* s_γ of a photon gas in ordinary three-space at temperature T is given by

$$s_\gamma = \frac{2}{45}\pi^2 T^3 \cdot 2\,, \tag{23.114}$$

where the last factor of two arises because the photon has two massless degrees of freedom. In the $SU(N)$ field theory the total entropy is given by

$$S_{YM} = \frac{2}{45}\pi^2 \bar{T}_H^3 \cdot \left(8 + 8 \cdot \frac{7}{8}\right) N^2 \cdot (2\pi^2 R^3)\,. \tag{23.115}$$

The last factor is the volume of the three-sphere of radius R where the field theory lives. In the limit of small 't Hooft coupling the massless particles in the $SU(N)$ theory are very weakly interacting and the total entropy is obtained by adding their separate contributions. The spectrum of the theory includes eight bosonic degrees of freedom and eight fermionic degrees of freedom, both repeated $N^2 - 1$ times, or just N^2 times, to the level of precision we need. Moreover, fermionic degrees of freedom contribute $7/8$ of the entropy of bosonic degrees of freedom. All these considerations explain the result given in (23.115). Simplifying this expression, we get

$$S_{YM} = \frac{2}{3}\pi^2 N^2 \bar{T}_H^3 \cdot (2\pi^2 R^3)\,, \quad \lambda \ll 1\,. \tag{23.116}$$

On the gravity side the entropy $S_{\rm BH}$ of the black hole is given by the "area" of the horizon divided by the five-dimensional Newton constant $G^{(5)}$:

$$S_{\rm BH} = \frac{A_{\rm hor}}{4G^{(5)}} \,. \tag{23.117}$$

The horizon is the locus of points with $r = r_+$. The last term in the metric (23.104) shows that the horizon is a three-sphere of radius r_+, so $A_{\rm hor} = 2\pi^2 r_+^3$. This, together with the value of $G^{(5)}$ from (23.63) and the relation of r_+ to $\bar{T}_{\rm H}$ in (23.111) gives

$$S_{\rm BH} = \frac{2\pi^2 r_+^3}{\frac{2\pi R^3}{N^2}} = \frac{N^2\, 2\pi^2 R^3\, r_+^3}{2\pi R^6} = \frac{1}{2}\pi^2 N^2 \bar{T}_{\rm H}^3 \cdot (2\pi^2 R^3) \,, \quad \lambda \gg 1 \,. \tag{23.118}$$

The assumption of large λ is implicit whenever we use the black hole picture. The result (23.118) is in qualitative agreement with the field theory estimate (23.116). The linear growth of the temperature with the radius of the black hole was essential to the match. Nevertheless, we got

$$S_{\rm BH} = \frac{3}{4} S_{\rm YM} \,, \tag{23.119}$$

so the agreement is not exact. It did not have to be so: in the field theory we assumed small 't Hooft parameter λ, while in the gravity side we assumed large λ. It is hoped that an exact calculation for arbitrary λ will show an entropy that decreases from the zero-coupling value obtained easily in the field theory to the large-coupling value obtained easily from the black hole.

There is an additional twist to the finite temperature correspondence. We have noted in (23.112) that, in a given AdS space with fixed radius of curvature R, all black holes have Hawking temperatures $\bar{T}_{\rm H} \geq T_0$ (see Figure 23.14). Consider the $SU(N)$ gauge theory at some strong coupling that is dual to AdS with radius of curvature R. For temperatures smaller than T_0 the field theory cannot have a black hole dual since there is no black hole. The dual is then just thermal AdS, an AdS space filled with a gas of particles at the specified temperature. The field theory at these low temperatures is in the analog of a *confined* phase: the entropy of the theory is of $\mathcal{O}(N^0)$. For temperatures greater than T_0 we have two black holes: any horizontal line above T_0 in Figure 23.14 cuts the curve at two values of r_+/R, one smaller than $1/\sqrt{2}$ and one larger than $1/\sqrt{2}$. Thermal AdS also remains a possible gravitational background. It turns out that for temperatures T in the range

$$T_0 \,\leq\, T \,\leq\, T_1 \,, \quad \text{with} \quad T_1 = \frac{3}{2\pi R} \,, \tag{23.120}$$

thermal AdS remains the thermodynamically favored state – it has lower free energy than either black hole background. The field theory remains confined for $T \leq T_1$. Note that T_1 is only a bit larger than T_0 (see Figure 23.14).

Quick calculation 23.10 What are the values of r_+/R for the two black holes that exist for temperature T_1? Compare with Figure 23.14.

As the temperature exceeds T_1 the thermodynamically favored background ceases to be thermal AdS and becomes the larger black hole – this is referred to as the Hawking–Page transition. In fact, the smaller black hole is always least favored among the three options; like black holes in Minkowski space it is unstable to evaporation. At the temperature T_1 the large-N, strongly coupled gauge theory goes from confined to *deconfined*: the entropy becomes of $\mathcal{O}(N^2)$, indicating liberation of elementary degrees of freedom. The dual of the deconfined gauge theory is the (large) AdS black hole.

This finite temperature discussion, where the black hole spacetime is not exactly anti-de Sitter, points to a more general statement of the AdS/CFT correspondence. The correspondence can now be stated as the equivalence of four-dimensional, supersymmetric $SU(N)$ gauge theory with type IIB superstrings in a space whose geometry is *asymptotically* $\mathrm{AdS}_5 \times S^5$. This statement clearly applies at zero and at finite temperatures. Further examples of asymptotically $\mathrm{AdS}_5 \times S^5$ spaces that match with particular states of the field theory have been found.

23.11 The quark–gluon plasma

Experiments at the Relativistic Heavy Ion Collider (RHIC) in Brookhaven National Laboratory have created, for very brief instants, a deconfined state of QCD. Physicists accelerated and collided nuclei of gold against each other. In the center of mass frame each nucleon packed an energy of about 100 GeV. With 197 nucleons, each gold nucleus at collision carried an energy of about 20 TeV. The collision appears to create a quark–gluon plasma (QGP), a strongly interacting system of deconfined quarks and gluons.

We pause to note the scales of energies and distances relevant to these collisions. Nuclei are smaller by a factor of about 100 000 than atoms, which have sizes of about 10^{-10}m. The natural unit to describe a nucleus is therefore the femtometer (fm): $1\,\mathrm{fm} = 10^{-15}$m. An approximate formula gives the radius r of a nucleus with A nucleons: $r \simeq r_0 A^{1/3}$, where $r_0 = 1.2\,\mathrm{fm}$. It is convenient to measure time in femtometers too, using the speed of light as a conversion unit. Thus, a time of 1fm is $\frac{1}{3} \times 10^{-23}$ s.

Quick calculation 23.11 Confirm that the estimate for the nuclear radius implies an energy density of inside nuclei of about $0.13\,\mathrm{GeV/fm}^3$. Confirm that the radius of the gold nucleus is about 7 fm.

Lattice calculations suggest that deconfinement in QCD occurs at a critical temperature $T_c \simeq 175\,\mathrm{MeV}$, with an uncertainty of about 10% due to systematic errors. This temperature corresponds to about two trillion degrees Kelvin, the temperature of the universe about 10^{-11} s after the Big Bang. At this temperature the energy density of the plasma is $0.7\,\mathrm{GeV/fm}^3$, about five to six times the nuclear energy density.

Since the colliding nuclei at RHIC have Lorentz factors $\gamma \sim 100$, in the center of mass frame they look like thin pancakes with thickness below $14\,\mathrm{fm}/100 = 0.14\,\mathrm{fm}$. The QGP

is created when the pancakes go through each other and is believed to attain thermal equilibrium at times as early as 1 fm. The QGP lasts for no more than 15 fm, at which time the quarks and gluons have recombined into hadrons that are later captured by the detectors. By measuring the energies of these hadrons it is estimated that the energy density in the QGP at time 1 fm is at least $5\,\mathrm{GeV/fm}^3$. For a head-on collision, at this time the plasma extends over a cylinder of length 2 fm and radius equal to the gold nucleus radius.

Quick calculation 23.12 Calculate the energy in the QGP for a head-on collision and show that it is about 4% of the energy available in the center-of-mass frame.

A temperature estimate is possible. Recalling that for a gas of massless particles the energy density $U \sim T^4$, we have

$$\left(\frac{T_{\mathrm{RHIC}}}{T_c}\right)^4 \simeq \frac{5\ \mathrm{GeV/fm}^3}{0.7\,\mathrm{GeV/fm}^3} \qquad \rightarrow \qquad T_{\mathrm{RHIC}} \simeq 1.6\,T_c\,. \tag{23.121}$$

This estimate suggests that at time equal to 1 fm the temperature is still comfortably above the deconfinement temperature. At the Large Hadron Collider (LHC) in CERN one may reach temperatures of about $5\,T_c$.

In general, collisions are non-central and the QGP is initially created in the approximate shape of an ellipse defined by the overlap of two "pancakes" with offset centers. The direction of the impact parameter is along the short axis of the ellipse. Consider the particles created by the collision with momenta orthogonal to the beam axis. One gets more particles along the direction of the short axis of the ellipse than along the direction of the long axis of the ellipse. Numerical simulations indicate that this anisotropy is consistent with the assumption that the QGP is a fluid with extraordinarily small viscosity. In a droplet of fluid one has maximal pressure at the center and zero pressure at the edges. The pressure gradient is larger along the shorter axis of the ellipse leading to a larger number of particles emitted in this direction.

It is natural to wonder if the QGP can be studied using the finite temperature version of the AdS/CFT correspondence. At first sight this seems unlikely since QCD, a non-supersymmetric $SU(3)$ gauge theory, is very different from supersymmetric $SU(N)$ gauge theory. Nevertheless there are some facts that make the proposition less outrageous. First, the QGP is strongly coupled: even with $N = 3$ one is likely to have a large 't Hooft coupling $\lambda \sim 20$, as is needed for a calculable gravity side. Moreover, at finite temperature supersymmetry is a broken symmetry and its effects are therefore somewhat hidden. It is plausible that finite temperature QCD and finite temperature supersymmetric gauge theory are not all that different.

It is striking that the general property of very small viscosity is a rather direct consequence of the AdS/CFT correspondence. The viscosity η of a fluid tells how the force F transmitted across fluid layers of area A depends on the velocity gradient $|\nabla v|$:

$$\frac{F}{A} = \eta |\nabla v|\,. \tag{23.122}$$

The units of viscosity are $[\eta] = M/(LT)$. The entropy density s is another important property of a fluid and has units $[s] = [k]/L^3$, where k is the Boltzmann constant. It follows that the η/s has units of $\hbar/k$. The AdS/CFT correspondence suggests that

$$\frac{\eta}{s} \geq \frac{\hbar}{4\pi k} \simeq 0.08 \cdot \frac{\hbar}{k} \,. \tag{23.123}$$

Indeed, at strong coupling, the gravitational calculation of η/s gives the result that saturates the above inequality. The viscosity η is related to the absorption cross section of the black hole and the entropy density is the one following from (23.118). Analysis of the data at RHIC is consistent with a value of η/s that is at most two or three times the bound. It seems likely that the bound (23.123) applies to any known fluid or gas in nature. For the case of a dilute gas the bound of η/s may be related to the Heisenberg uncertainty principle (Problem 23.8).

Quick calculation 23.13 At 25 °C and a pressure of 1 atm, liquid water has a viscosity of 0.9×10^{-3} kg/(m · s) and a molar entropy of 70 J/K. Calculate the ratio η/s and confirm that it exceeds the bound by a factor of nearly four hundred. (Useful constant: $k = 1.38 \times 10^{-23}$ J/K.)

Another interesting discovery at RHIC is that very high energy quarks (jets) moving through the QGP are stopped, or "quenched," after traveling only a few femtometers. This property of the QCD plasma is parameterized by a jet quenching parameter $\hat{q}$. The energy loss ΔE of the jet is proportional to $\hat{q}$ and, quite strikingly, to the *square* of the distance L travelled in the plasma: $\Delta E \sim \hat{q} L^2$. The above indicates that $\hat{q}$ has units of energy over length-squared. In natural units ($\hbar = c = 1$) this is equivalent to energy-squared over length, and this is the way that $\hat{q}$ is traditionally presented. The experimental value of $\hat{q}$ is fairly uncertain since the QGP expands and cools and one only sees the effects of a time-varying $\hat{q}$. At time 1 fm the value of $\hat{q}$ appears to lie in the range $5 - 15\,\mathrm{GeV}^2/\mathrm{fm}$. The parameter $\hat{q}$ of QCD can be given a natural gauge theory definition whose value in strongly coupled, finite temperature, supersymmetric Yang–Mills can be calculated using the dual black hole background. This result can be estimated at time 1 fm, using a 't Hooft coupling $\lambda \sim 20$, and the result is $\hat{q} \sim 4\,\mathrm{GeV}^2/\mathrm{fm}$. Although the analytic result for super-Yang–Mills gives a $\hat{q}$ lower than the one for hot QCD, given the various theoretical and experimental uncertainties this level of coincidence is encouraging. It seems likely that the AdS/CFT will be a valuable tool to understand and perhaps predict properties of the QGP to be measured at the higher energies of the LHC.

Problems

Problem 23.1 Length of a rotating string.

Consider the normalized string state that represents an open string rotating on the (x^2, x^3) plane:

$$|\psi_N\rangle = \frac{1}{\sqrt{N!}} (\alpha_{-1})^N |p^+, \vec{0}\rangle \,. \tag{1}$$

As discussed above (23.17), we write $\langle p^+, \vec{0}|p^+, \vec{0}\rangle = 1$, instead of the correct but cumbersome delta function of zero argument. To estimate the length of the string state we will evaluate the expectation value of a length-squared operator L^2 defined by

$$L^2(\tau) \equiv : (\Delta X^2(\tau)\Delta X^2(\tau) + \Delta X^3(\tau)\Delta X^3(\tau)) : . \tag{2}$$

Here $\Delta X^I(\tau) = X^I(\tau, \pi) - X^I(\tau, 0)$ is the difference of X^I coordinates at the string endpoints and the colons denote normal ordering. Given a product of oscillators that includes both creation and annihilation operators, normal ordering places all annihilators to the right of the creators. Thus, for example, $:\alpha_1^i \alpha_{-1}^j: = \alpha_{-1}^j \alpha_1^i$ and normal ordering does not affect $\alpha_{-1}^i \alpha_{-1}^j$ nor $\alpha_1^i \alpha_1^j$. Normal ordering is needed in (2) because otherwise L^2 would have infinite expectation value even on the ground states. It is useful to define string coordinates

$$X \equiv \frac{1}{\sqrt{2}}(X^I + iX^2)\,, \qquad \bar{X} \equiv \frac{1}{\sqrt{2}}(X^I - iX^2)\,.$$

Verify that $L^2(\tau) =: 2\Delta X \Delta \bar{X} :$ and use this expression to calculate the expectation value $\langle L^2\rangle = \langle \psi_N | L^2 | \psi_N \rangle$. Show that for states with large N, $\sqrt{\langle L^2\rangle}$ gives the expected classical value of the length of the string. To test further your result, verify that a rotating string of $M^2 = 3/\alpha'$ has a length $\sqrt{\langle L^2\rangle}$ equal to eight times the string length. [Hint: the expanded out $:2\Delta X \Delta \bar{X}:$ contains four kinds of terms, each with a different structure of α and $\bar{\alpha}$ oscillators. Only one contributes to the expectation value.]

Problem 23.2 Coherent states for quantum oscillator.

In order to understand better (23.18) we examine analogous coherent states

$$|z\rangle \equiv e^{za^\dagger - z^* a}|0\rangle\,,$$

of the simple harmonic oscillator. Here z^* is the complex conjugate of the complex number z. For the oscillator we use the Hamiltonian $H = a^\dagger a$ with $a = \frac{1}{\sqrt{2}}(x + ip)$.

(a) Explain why $\langle z|z\rangle = 1$. Show that $|z\rangle = e^{-\frac{1}{2}|z|^2} e^{za^\dagger}|0\rangle$.
(b) Show that the expectation values of x and p in the state $|z\rangle$ are encoded in the real and imaginary parts of z via the relation $\frac{1}{\sqrt{2}}\big(\langle x\rangle + i\langle p\rangle\big) = z$.
(c) Calculate the time-dependent physical state $e^{-iHt}|z_0\rangle$, where z_0 is an arbitrary complex number. Find the corresponding time-dependent expectations values of x and p in terms of the constants x_0 and p_0 defined by $z_0 \equiv \frac{1}{\sqrt{2}}(x_0 + ip_0)$.
(d) Confirm that the time-dependent expectation values of x and p satisfy the classical equations of motion of the oscillator.

Problem 23.3 Isometric embedding of a portion of $\mathbb{H}_2$ in $\mathbb{R}^3$ using the *tractrix*.

The metric (23.77) on $\mathbb{H}_2$ can be neatly written using a complex variable $w = \xi^1 + i\xi^2$:

$$ds^2 = R^2 \frac{4\,(d\xi^1 d\xi^1 + d\xi^2 d\xi^2)}{(1 - (\xi^1\xi^1 + \xi^2\xi^2))^2} = R^2 \frac{4\,dw d\bar{w}}{(1 - w\bar{w})^2}\,. \tag{1}$$

The complex variable w is constrained to lie within the unit disk: $|w| < 1$.

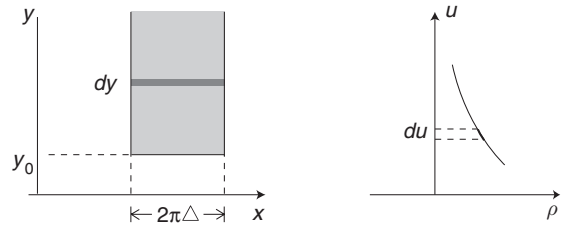

Fig. 23.15 **Problem 23.3. Left: a portion of $\mathbb{H}_2$ to be isometrically embedded in $\mathbb{R}^3$. Right: The curve $u(\rho)$, rotated about the u axis, gives the embedded surface.**

(a) It is more convenient to describe $\mathbb{H}_2$ using the upper-half plane (UHP), the part of the complex plane above the real line. If we use a complex coordinate $z = x + iy$ to describe the complex plane, the conformal map

$$z = \frac{1}{i}\,\frac{w-1}{w+1}\,, \tag{2}$$

takes the region $|w| < 1$ to the region $\Im(z) > 0$ of the complex z-plane. To convince yourself of this fact, show that the circle $w = e^{i\theta}$ is mapped by (2) to the full real line and give the resulting function $x(\theta)$. [Hint: the answer is a simple function of $\theta/2$.] Then verify that one point in the interior of the w-disk goes to a point above the real axis on the z-plane.

(b) Prove that in z coordinates the metric (1) takes the form

$$ds^2 = R^2\,\frac{dz\,d\bar{z}}{(\Im(z))^2} = R^2\,\frac{dx^2+dy^2}{y^2}\,. \tag{3}$$

It is possible to embed isometrically in $\mathbb{R}^3$ a *piece* of the $\mathbb{H}_2$ surface. Consider Figure 23.15, where we show a semi-infinite vertical strip of coordinate width $\Delta x = 2\pi\,\Delta$ for some arbitrarily chosen positive number Δ. As we will see, the embedding of the strip works down to $y = y_0$, where $y_0 > 0$ depends on Δ. It is convenient to turn the strip into a cylinder by identifying the vertical edges. This identification is consistent with the metric, which is manifestly invariant under horizontal translations $x \to x + c,\ \ y \to y$. The embedded surface is the surface of revolution obtained by spinning the curve $u(\rho)$, shown to the right of the figure, about the u axis.

(c) Assume that the embedding is done with u increasing as y increases. It follows that ρ decreases with increasing u (why?). Consider the little circular strip of coordinate width dy shown to the left of the figure and the corresponding one of height du to the right of the figure, after spinning the curve. By equating the length and areas of these pieces of surface derive the following differential equation for $u(\rho)$:

$$\frac{du}{d\rho} = -\frac{\sqrt{R^2-\rho^2}}{\rho}\,. \tag{4}$$

Clearly, $\rho \le R$. Find the relation between y_0 and Δ. The curve $u(\rho)$ is a *tractrix*.

(d) For any point P on the curve $u(\rho)$ consider the point P' on the u axis such that the segment PP' is tangent to the curve at P. Use the differential equation (4) (not its solution!) to show that the distance between P and P' is always equal to R. This property was used by Newton to define the tractrix. You can easily construct a tractrix as follows. Attach to a small heavy object a piece of string of length R and place the object on a table, a distance R away from a straight edge. If you drag the object by sliding the end of the string along the edge of the table, the object will move along the tractrix.

(e) Integrate (4) using $u(R) = 0$ and give the explicit form of $u(\rho)$. If you get the correct answer, $\rho = \frac{3}{5}R$ should give $u = R(-\frac{4}{5} + \ln 3)$. Make a plot of $u(\rho)$.

Problem 23.4 A conformal view of two-dimensional Minkowski space.

The metric of two-dimensional Minkowski space is $-ds^2 = -dt^2 + dx^2$, where the coordinates (x, t) take all possible real values: $-\infty < t, x < \infty$. Consider new coordinates (x', t') defined by the two relations

$$\tan(x' \pm t') = x \pm t\,.$$

(a) Show that in the new coordinates the metric takes the form

$$-ds^2 = \Omega^2(t', x')(-dt'^2 + dx'^2)\,,$$

and determine the function Ω^2. Show that the full (x, t) plane is mapped to a finite region of the (x', t') plane. Describe this region and sketch it.

(b) Consider light-rays that depart from $x = 0$ at times $t = -1, 0$, and 1, and move towards x positive. Sketch the three light-rays in the (x', t') plane.

(c) Consider the timelike line $x = \alpha t$ with α a positive constant that satisfies $\alpha < 1$ and $t \in [0, \infty)$. Show that in the (x', t') plane the corresponding trajectory starts at the origin and eventually approaches the point $(x', t') = (0, \pi/2)$. Moreover, near that point, the trajectory follows the equation $t' = \frac{\pi}{2} - \frac{x'}{\alpha}$. Discuss the corresponding results for spacelike lines $x = \alpha t$ with $\alpha > 1$ and $t \in [0, \infty)$. You may find the following large u expansions useful:

$$\tan^{-1} u \simeq \pm\frac{\pi}{2} - \frac{1}{u}\,, \quad u \to \pm\infty\,.$$

Problem 23.5 Another view of the full AdS_{n+1} space.

(a) Consider the ambient metric and constraint (23.9) and now set $z = R\cosh\rho$, where $\rho \in [0, \infty)$ is a new coordinate. Moreover, solve the constraint by letting

$$x^i = R\,\Omega^i \sinh\rho \quad \text{with} \quad \Omega^i\Omega^i = 1\,. \tag{1}$$

The constrained variables Ω^i are coordinates on the unit sphere S^{n-1}. Note that $\Omega^i\Omega^i = 1$ requires the differentials to satisfy $\Omega^i d\Omega^i = 0$. Show that the metric becomes

$$ds^2 = R^2\big[-(\cosh\rho)^2\,(dt)^2 + (d\rho)^2 + (\sinh\rho)^2 d\Omega_{n-1}^2\big]\,, \tag{2}$$

where $d\Omega_{n-1}^2 \equiv d\Omega^i d\Omega^i$, given the constraint $\Omega^i \Omega^i = 1$, is the metric on the unit sphere S^{n-1}. Equation (2) is a useful form of the AdS_{n+1} metric.

(b) Consider the n-dimensional spatial region $\rho \le \bar{\rho}$. Write an integral that gives its volume $V(\bar{\rho})$ and an expression for the area $A(\bar{\rho})$ of its boundary. Show that

$$\lim_{\rho\to\infty} \frac{A(\bar{\rho})}{V(\bar{\rho})} = \frac{c_n}{R}\,, \tag{3}$$

where the constant c_n that you will determine is consistent with (23.99). [Hint: the leading divergent term in the integral is easy to compute, the rest does not contribute in the limit.]

(c) In (2) introduce a new radial variable $r \in [0, \infty)$ by the relation $r = R \sinh \rho$. Also let $t \to t/R$. Show that the metric becomes

$$ds^2 = -\Big(1 + \frac{r^2}{R^2}\Big)(dt)^2 + \Big(1 + \frac{r^2}{R^2}\Big)^{-1}(dr)^2 + r^2 d\Omega_{n-1}^2\,. \tag{4}$$

This is another useful form of the AdS_{n+1} metric. The simple modification indicated in (23.104) gives the metric of the Schwarzschild black hole with asymptotically AdS geometry.

Problem 23.6 A partial but simple view of AdS_{n+1} space.

Consider the ambient metric and constraint (23.84) written as

$$\text{ambient metric:} \quad ds^2 = -(du)^2 + \sum_{i=1}^{n-1} dx^i dx^i + (dx^n)^2 - (dv)^2\,,$$

$$\text{constraint:} \quad -u^2 + \sum_{i=1}^{n-1} x^i x^i + (x^n)^2 - v^2 = -R^2\,.$$

(a) We now introduce coordinates z, t, and $\vec{y} = (y^1, \ldots, y^{n-1})$ by the relations:

$$v + x^n = \frac{R}{z}\,, \quad u = \frac{R}{z}\,t\,, \quad x^i = \frac{R}{z}\,y^i\,,\ i = 1, \ldots, n-1\,. \tag{1}$$

Here $t \in \mathbb{R}$, $\vec{y} \in \mathbb{R}^{n-1}$, and $z \in \mathbb{R}^+$. Since z is positive, so is $x^n + v$ and this presentation will not display the part of AdS space where $x^n + v$ is negative. Use the constraint to determine the value of $v - x^n$ in terms of the new coordinates.

(b) Show that the metric on the AdS_{n+1} space takes the remarkably simple form

$$ds^2 = \frac{R^2}{z^2}\Big(dz^2 - (dt)^2 + (d\vec{y})^2\Big)\,. \tag{2}$$

[Hint: for the evaluation use $(dx^n)^2 - (dv)^2 = d(x^n + v)d(x^n - v)$.] The boundary of the space is at $z = 0$, where the metric is conformal to $-(dt)^2 + (d\vec{y})^2$, the n-dimensional Minkowski metric. A related form is obtained letting $z = R^2/r$:

$$ds^2 = R^2\,\frac{dr^2}{r^2} + \frac{r^2}{R^2}\big[-(dt)^2 + (d\vec{y})^2\big]\,. \tag{3}$$

Problem 23.7 De Sitter spacetime.

The n-dimensional de Sitter spacetime dS_n has positive curvature. Its definition is a variant of that for hyperbolic spaces: one can obtain de Sitter spacetime as a surface in a Minkowski space of one higher dimension. This time, however, the normal to the surface is spacelike so that the surface contains a time direction. Moreover, just like hyperbolic spaces can describe spatial sections of AdS, spheres can be used to describe the spatial sections of dS.

With $i = 1, 2, \ldots, n$, the dS_n spacetime and its metric are defined by

$$\begin{aligned} \text{ambient metric:} \quad & ds^2 = -(dz)^2 + dx^i dx^i \,, \\ \text{constraint:} \quad & -z^2 + x^i x^i = R^2 \,. \end{aligned} \tag{1}$$

(a) Explain in detail why the surface contains a timelike direction. Exhibit the timelike tangent at points on the surface with $z \neq 0$ and at points with $z = 0$.

(b) To solve the constraint we set

$$z = R \sinh t \,, \quad x^i = R\,\Omega^i \cosh t \,, \quad \text{with} \quad \Omega^i \Omega^i = 1 \,. \tag{2}$$

See Problem 23.5 for comments on the constrained variables Ω^i. Show that the metric on the dS_n space takes the form

$$ds^2 = R^2\big[-(dt)^2 + (\cosh t)^2 d\Omega_{n-1}^2\big] \,, \tag{3}$$

where $d\Omega_{n-1}^2 \equiv d\Omega^i d\Omega^i$ is the metric on the unit sphere S^{n-1}. At any fixed time the spatial section is a sphere of radius $R \cosh t$. As time increases, the spatial section contracts while $t < 0$ and expands while $t > 0$.

(c) We noted in Section 23.8 that the mass-shell for a massive particle is hyperbolic space. What is the mass-shell for a tachyon?

Problem 23.8 Viscosity and entropy of a dilute gas.

Consider a container of volume V with N molecules of a gas. The gas molecules have mass m, an average velocity $\bar{v}$, and their mean free path is ℓ. Assuming that the gas is dilute, the viscosity of the gas is given roughly by

$$\eta = \frac{1}{3}\frac{N}{V}\,\bar{v} m \ell \,. \tag{1}$$

For a monoatomic ideal gas the entropy density s given by

$$\frac{s}{k} = \frac{1}{k}\frac{S}{V} = \frac{N}{V}\left[\ln\Big(\frac{V/N}{\lambda_{\text{th}}^3}\Big) + \frac{5}{2}\right], \quad \lambda_{\text{th}} = \frac{h}{\sqrt{2\pi m k T}} \,. \tag{2}$$

Here λ_{th} is the thermal de Broglie wavelength. You can find the derivation of the above results for η and s in many textbooks on thermal physics.

(a) Use (2) to calculate the value of s/k per unit cm^3 for monoatomic hydrogen at $T = 25\,^\circ\text{C}$ and 1 atm. Compare with the molar entropy of 114.7 J/K quoted by chemical tables.

(b) Consult a textbook to find the Bose–Einstein condensation temperature T_{B}. Show that at this temperature the factor in brackets in (2) takes the approximate value 1.5397. So, as long as we have a gas, $s/k > (1.5397)\,\frac{N}{V}$. Show that for temperatures below one-thousand times T_{B}, we have $s/k < 11.901\,\frac{N}{V}$.

(c) Estimate the mean free path ℓ in terms of N, V, and the molecular diameter d. Recalling that $\bar{v} \sim \sqrt{kT/m}$ conclude that $\eta \sim T^{1/2}$ and is independent of the pressure or the density. For a gas the viscosity increases with the temperature. For a liquid the viscosity decreases with temperature.

(d) In a gas $(\bar{v})^2 \simeq \langle v^2 \rangle$. Use this to show that $\eta = \frac{2}{3}\frac{N}{V} u\tau$, where u is the average energy of a molecule and τ is the mean free time. Conclude that in the range of temperatures considered in (b) one has

$$\frac{\eta}{s} \geq \frac{\eta}{s_{max}} \simeq 0.056\,\frac{1}{k}\,u\,\tau\,. \tag{3}$$

The energy/time form of the uncertainty principle suggests that $u\tau \geq \hbar$, leading to η/s bounded below by some fraction of $\hbar/k$ (compare with (23.123)). The estimate (3) is rough; it does not incorporate the growth of η over the temperature range.

24 Covariant string quantization

In the Lorentz covariant quantization of string theory we treat all string coordinates $X^\mu(\tau, \sigma)$ on the same footing. To select physical states we use the constraints generated by a subset of the Virasoro operators. The states automatically carry time labels, so the Hamiltonian does not generate time evolution. We describe the Polyakov string action and show that it is classically equivalent to the Nambu–Goto action.

24.1 Introduction

In this book, the quantization of strings was carried out using light-cone coordinates and the light-cone gauge. String theory is a Lorentz invariant theory, but Lorentz symmetry is not manifest in the light-cone quantum theory. Indeed, the choice of a particular coordinate X^+ for special treatment hides from plain view the Lorentz symmetry of the theory. While hidden, the Lorentz symmetry is still a symmetry of the quantum theory, as we demonstrated by the construction of the Lorentz generator M^{-I}. This generator has the expected properties when the spacetime has the critical dimension.

Since Lorentz symmetry is of central importance, it is natural to ask if we can quantize strings preserving *manifest* Lorentz invariance. It is indeed possible to do so. The Lorentz covariant quantization has some advantages over the light-cone quantization. Our light-cone quantization of open strings did not apply to D0-branes because the light-cone gauge requires that at least one spatial open string coordinate has Neumann boundary conditions. Covariant quantization applies to D0-branes. The equations of motion for the fields that arise in string theory are better understood in Lorentz covariant notation. The calculation of tachyon potentials, alluded to in Section 12.8, appears to be possible only within the Lorentz covariant quantization of strings.

Why then, have we waited so long to discuss the Lorentz covariant quantization of strings? The covariant approach is very elegant but it is sometimes hard to extract the physical content from its equations. Moreover, covariant quantization has a series of features that are quite strange. We are accustomed to the idea that in quantum mechanics the position of a particle becomes an operator while time remains a parameter. In a Lorentz invariant quantization, all the coordinates x^μ of a particle, including x^0, become operators. Similar remarks apply to the string coordinates X^μ. We will also see that the string Hamiltonian annihilates the physical states of the theory. Finally, in covariant quantization it is

necessary to discuss states whose norm is not positive; this takes us out of the usual Hilbert space postulates.

We chose to do the quantization of strings in the light-cone gauge because all of the above features would have distracted us from the task of extracting the physical content of the theory. The light-cone gauge has served us well, and it will continue to do so in the following chapters, where we use light-cone string diagrams to begin our discussion of string interactions. The proper treatment of covariant quantization requires tools that go beyond the level of this book. We will not be able to derive the critical dimension, for example. While our treatment will not be complete, we will still gain some important insights into the structure of the theory.

Let us begin our discussion by recalling some facts about parameterizations of the world-sheet. In Chapter 9 we described a large class of gauges characterized by a vector n^μ. By choosing n^μ appropriately we could produce the static or light-cone gauge. For open strings the choice of n^μ completely fixes the parameterization of the world-sheet. This is almost true for closed strings as well, except that we are still free to shift the coordinate σ rigidly along the strings. We showed that for any choice of n^μ the string coordinates satisfy the constraints

$$(\dot{X} \pm X')^2 = 0\,. \tag{24.1}$$

Since *any* choice of n^μ results in these constraints, the constraints by themselves do not completely fix the parameterization of the world-sheet. In fact, as you may have seen in Problem 12.10, many reparameterizations preserve these constraints.

We were able to show that the constraints (24.1) cause the equations of motion to become simple wave equations:

$$\ddot{X}^\mu - X^{\mu\prime\prime} = 0\,. \tag{24.2}$$

Furthermore, they imply that the momentum densities are given by

$$\mathcal{P}^{\sigma\mu} = -\frac{1}{2\pi\alpha'}\,X'^\mu\,, \quad \mathcal{P}^{\tau\mu} = \frac{1}{2\pi\alpha'}\,\dot{X}^\mu\,. \tag{24.3}$$

In the covariant formalism we use the constraints (24.1), and thus also (24.2) and (24.3), but we do *not* fix completely the parameterization of the world-sheet. The constraints (24.1) can be thought of as conditions for a *partial* gauge fixing. At the classical level, we solve the wave equations and check that the constraints are satisfied. At the quantum level, the constraints introduce subtle complications.

As you probably recall, we had to do plenty of work to get wave equations and simple momentum densities from the Nambu–Goto action. Is there an action for which these results follow quickly? The answer is yes. In fact, for transverse light-cone coordinates, such an action was given in (12.81). We used this action to give a more physical derivation of the oscillator commutation relations. In the present case, the desired action is

$$S = \int d\tau d\sigma\,\mathcal{L} = \frac{1}{4\pi\alpha'}\int d\tau d\sigma\,\left(\partial_\tau X^\mu \partial_\tau X_\mu - \partial_\sigma X^\mu \partial_\sigma X_\mu\right). \tag{24.4}$$

Note that this action, as opposed to the Nambu–Goto action, contains no square root. It is quadratic in the dynamical variables X^μ. This action is useful because variation of X^μ

immediately produces the wave equations (24.2). Moreover, the simple expressions for the canonical momentum densities arise directly. For example,

$$\mathcal{P}_\mu^\tau = \frac{\partial \mathcal{L}}{\partial \dot{X}^\mu} = \frac{1}{2\pi\alpha'}\dot{X}_\mu \,. \tag{24.5}$$

Even the classical Hamiltonian is easy to calculate. In terms of coordinates and momenta, and dropping the τ superscript from $\mathcal{P}_\mu^\tau$, we find

$$\begin{aligned} H &= \int d\sigma\,(\,\mathcal{P}_\mu\,\dot{X}^\mu - \mathcal{L}\,) \\ &= \int d\sigma \left(\mathcal{P}_\mu \dot{X}^\mu - \frac{1}{4\pi\alpha'}\left((2\pi\alpha'\mathcal{P})^2 - X'^2\right)\right) \\ &= \pi\alpha' \int d\sigma \left(\mathcal{P}\cdot\mathcal{P} + \frac{X'\cdot X'}{(2\pi\alpha')^2}\right). \end{aligned} \tag{24.6}$$

This Hamiltonian is analogous to the light-cone Hamiltonian (12.15), but here the index contractions run over all spacetime dimensions. In the light-cone quantization we defined Heisenberg operators $X^I(\tau,\sigma)$ and $\mathcal{P}_I(\tau,\sigma)$ which had canonical commutation relations. Using these we verified that the Hamiltonian generates the correct operator equations of motion. In the covariant theory it is natural to introduce Heisenberg operators

$$X^\mu(\tau,\sigma) \quad \text{and} \quad \mathcal{P}_\mu(\tau,\sigma) \tag{24.7}$$

and to postulate the commutation relations

$$\left[X^\mu(\tau,\sigma)\,,\ \mathcal{P}^\nu(\tau,\sigma')\right] = i\,\eta^{\mu\nu}\delta(\sigma-\sigma')\,. \tag{24.8}$$

Note that even X^0 is a quantum operator. As usual, we set to zero the commutator of coordinates with coordinates and the commutator of momenta with momenta. Computations similar to those we performed using the light-cone gauge show that the quantum equations of motion take the form (24.2). This is strong evidence that H is the correct Hamiltonian.

24.2 Open string Virasoro operators

In light-cone quantization the constraints were used to solve for X^- in terms of the transverse coordinates X^I. The modes of the X^- coordinate were identified as transverse Virasoro operators. In the covariant approach the quantum constraints are not solved; rather, they are imposed on the states of the theory. We now examine the quantum constraints for the open string.

The quantization of the generic open string coordinate X^μ is similar to the quantization of the transverse light-cone coordinate X^I. Recall the oscillator expansion (9.56):

$$X^\mu(\tau,\sigma) = x_0^\mu + \sqrt{2\alpha'}\,\alpha_0^\mu\tau + i\sqrt{2\alpha'}\sum_{n\neq 0}\frac{1}{n}\alpha_n^\mu e^{-in\tau}\cos n\sigma\,. \tag{24.9}$$

This expansion led to particularly simple expressions for the linear combinations of derivatives in (9.59):

$$\dot{X}^\mu \pm X^{\mu\prime} = \sqrt{2\alpha'} \sum_{n\in\mathbb{Z}} \alpha_n^\mu \, e^{-in(\tau\pm\sigma)} \,. \tag{24.10}$$

This time there are no gauge conditions that we can use to simplify particular coordinates. The above expansions are to be used for all the string coordinates. The commutation relations (24.8), the expansions (24.10), and equation (24.5) together determine the oscillator commutation relations. Since the computations are analogous to the ones we performed in the light-cone gauge, the earlier results (12.45) and (12.64) are modified only by the replacement of η^{IJ} with $\eta^{\mu\nu}$:

$$[\,\alpha_m^\mu\,,\alpha_n^\nu\,] = m\,\eta^{\mu\nu}\delta_{m+n,0}\,, \quad [\,a_m^\mu\,,a_n^{\nu\dagger}\,] = \delta_{m,n}\,\eta^{\mu\nu}. \tag{24.11}$$

This time $\alpha_0^\mu = \sqrt{2\alpha'}\,p^\mu$, and the zero mode operators satisfy

$$[\,x^\mu\,,\;p^\nu\,] = i\eta^{\mu\nu}\,. \tag{24.12}$$

This equation should give us pause. We are familiar with the quantum mechanical procedure by which the spatial coordinates and the spatial momenta of a particle are turned into operators. In that scheme the spatial dependence of states is built-in, and the time dependence is determined by the Schrödinger equation. In covariant quantization even the time coordinate x^0 is made into an operator. So the states have a built-in time dependence. This means that the role of the Schrödinger equation must change. We will discuss this matter in detail in Section 24.4.

We can now explore the constraints (24.1) explicitly. Again, the computations are analogous to those of Chapter 9. By comparing with (9.79) we find

$$(\dot{X}\pm X')^2 = 4\alpha' \sum_{n\in\mathbb{Z}} L_n\, e^{-in(\tau\pm\sigma)}\,, \quad L_n = \frac{1}{2}\sum_{p\in\mathbb{Z}} \alpha_{n-p}^\mu\,\alpha_{p,\mu}\,. \tag{24.13}$$

The covariant Virasoro operators L_n differ from the transverse Virasoro operators $L_n^\perp$ because they include contributions from all the string coordinates. As before, the only Virasoro operator which has an ambiguous ordering is L_0. Again, we define L_0 to be the normal-ordered operator without any additional constant.

At the classical level, the constraint equations $(\dot{X}\pm X')^2 = 0$ require

$$\text{classically:} \quad L_n = 0\,, \quad n\in\mathbb{Z}\,. \tag{24.14}$$

Classically, we have $L_n^* = L_{-n}$, so that the constraints in (24.14) need only be checked for $n \geq 0$. At the quantum level we have $L_n^\dagger = L_{-n}$. The analogous property for the transverse Virasoro operators was proven below (12.112).

Quick calculation 24.1 Verify that, classically, $L_n^* = L_{-n}$.

The quantum Virasoro operators have rather nontrivial commutation relations. We calculated these commutators for the transverse Virasoro operators, and the result was given in (12.133). The central term was proportional to $D-2$, the number of transverse coordinates of string theory. Each transverse coordinate contributed the same amount to the central term. In the covariant treatment, where the Virasoro operators involve a sum over the full set of spacetime coordinates, that result becomes

$$[L_m, L_n] = (m-n)L_{m+n} + \frac{D}{12}(m^3 - m)\delta_{m+n,0}\,, \tag{24.15}$$

where $D = 26$ is the full dimensionality of spacetime (we will not derive the critical dimension here). You may be wondering why the time coordinate, whose oscillators commute to give a minus sign (see (24.11)), contributes to the central term in the same way as the space coordinates. It does so because the central term arises from a commutator which involves four oscillators, so the basic commutator is used twice and the sign cancels out.

In (12.139) we established that the action of the transverse Virasoro operators on the string coordinates X^I describes the effects of a certain class of world-sheet reparameterizations. Exactly the same results hold for the action of the covariant Virasoro operators on the string coordinates X^μ. The reparameterizations have a special property: they preserve the constraints $(\dot{X} \pm X')^2 = 0$, as shown in Problem 12.10. This result gives a physical interpretation to the covariant Virasoro operators. These operators generate world-sheet reparameterizations that preserve the gauge conditions of the Lorentz-covariant formalism.

24.3 Selecting the quantum constraints

The intuition inspired by the classical theory suggests that the physical states of the quantum theory – the quantum analog of consistent classical string motions – should be annihilated by all the Virasoro operators. But we will see that no states would survive the imposition of all these Virasoro constraints.

It is absolutely clear, however, that some constraints must be imposed. In covariant quantization we have 26 sets of oscillators available, so if we did not impose any constraint the set of quantum states would be different from the set of quantum states we obtained earlier, when we had only 24 sets of oscillators. Different quantization procedures should not give different physical results. It is therefore reasonable to expect that at least some Virasoro operators must annihilate the physical states.

Let us start by exploring the constraint that the Virasoro operator L_0 would impose on states. We have *defined* the quantum L_0 as normal-ordered with no additional constant, but there is no reason why this L_0 should annihilate physical states. Rather, we expect that $(L_0 + a)$, where a is some constant, should annihilate physical states. We investigated a similar issue in the light-cone gauge. The light-cone energy p^- turned out to be proportional to $L_0^\perp - 1$ rather than proportional to $L_0^\perp$ (see (12.157)). The same constant is needed in the covariant formalism, so the quantum constraint is

$$(L_0 - 1)|\Phi\rangle = 0\,, \tag{24.16}$$

for any physical state $|\Phi\rangle$. While we will not derive the ordering constant from first principles, we will see below that it is necessary for the covariant spectrum to agree with the one we obtained using the light-cone gauge. In the light-cone quantization, the relation between p^- and $L_0^\perp$ fixed the mass-squared of the states. The mass-squared is determined in the covariant treatment by the constraint (24.16). Writing the L_0 operator explicitly, we have

$$L_0 - 1 = \frac{1}{2}\alpha_0^\mu \alpha_{0,\mu} + \sum_{p=1}^{\infty} \alpha_{-p}^\mu \, \alpha_{p,\mu} \; - 1 = \alpha' p^2 - 1 + \sum_{n=1}^{\infty} n \, a_n^{\mu\dagger} \, a_{n,\mu} = 0 \,. \quad (24.17)$$

Since $M^2 = -p^2$, we find

$$M^2 = \frac{1}{\alpha'}\Big(-1 + N\Big) \,, \quad N = \sum_{n=1}^{\infty} n \, a_n^{\mu\dagger} \, a_{n,\mu} \,. \quad (24.18)$$

This result takes the same form as the light-cone result, except that the number operator $N^\perp$ is replaced by its covariant counterpart N. The eigenvalues of the light-cone number operator $N^\perp$ were manifestly non-negative, and this allowed us to conclude that $M^2 \geq -1/\alpha'$. Are the eigenvalues of the number operator N also non-negative? To find out, we expand out the sum over μ in the formula for N. With due attention to the Minkowski metric, we find

$$N = \sum_{n=1}^{\infty} n \left(-a_{n,0}^\dagger \, a_{n,0} + \sum_{i=1}^{25} a_{n,i}^\dagger \, a_{n,i} \right) . \quad (24.19)$$

In the *classical* theory the dagger represents complex conjugation, so the number operator N could be negative with appropriate contributions from the time components. In the quantum theory the situation improves. Because the commutator of two timelike oscillators carries a negative sign ($[a_{n,0}, a_{n,0}^\dagger] = -1$), we have

$$[\, -n \, a_{n,0}^\dagger \, a_{n,0} \,, \; a_{n,0}^\dagger \,] = +n \, a_{n,0}^\dagger \,. \quad (24.20)$$

So even the timelike oscillators contribute positively to N. Thus N is non-negative, and $M^2 \geq -1/\alpha'$ continues to hold in covariant quantization. Even more, ground states in the covariant theory have $N = 0$, and therefore $M^2 = -1/\alpha'$. This value agrees with the mass-squared of the ground states that we obtained using the light-cone quantization. The agreement justifies our selection of the ordering constant in (24.16).

There is a problem, however, which indicates that additional constraints must be imposed: many states do not have positive norm! Consider an eigenstate $|\Phi\rangle$ of N with eigenvalue N_0 and positive norm $\langle\Phi|\Phi\rangle > 0$. We then have $a_{n,0}|\Phi\rangle = 0$ for $n > N_0$. Now consider the state $|\chi\rangle = a_{n,0}^\dagger|\Phi\rangle$, with Hermitian conjugate $\langle\chi| = \langle\Phi|a_{n,0}$. The norm of this state is

$$\langle\chi|\chi\rangle = \langle\Phi|a_{n,0}\, a_{n,0}^\dagger|\Phi\rangle = \langle\Phi|\,[\, a_{n,0} \,, a_{n,0}^\dagger \,]\,|\Phi\rangle = -\langle\Phi|\Phi\rangle < 0 \,. \quad (24.21)$$

Since $|\chi\rangle$ has negative norm, it is an unacceptable quantum state. We must impose extra conditions to remove such physically impossible states from the spectrum. Recall that in the light-cone gauge quantization all states have positive norm.

Let us therefore attempt to establish the full set of constraints. We cannot require that, in addition to $L_0 - 1$, all other Virasoro operators annihilate the physical states. We show this by considering a simple example. We will prove that there are no nontrivial states annihilated by the three operators $L_0 - 1$, L_2 and L_{-2}. Consider therefore a general state $|\Phi\rangle$, and impose the conditions

$$(L_0 - 1)|\Phi\rangle = 0\,, \quad L_2|\Phi\rangle = 0\,, \quad \text{and} \quad L_{-2}|\Phi\rangle = 0\,.$$

If these equations hold, then the commutator of L_2 with L_{-2} must annihilate the state:

$$[L_2, L_{-2}]|\Phi\rangle = L_2(L_{-2}|\Phi\rangle) - L_{-2}(L_2|\Phi\rangle) = 0\,. \tag{24.22}$$

On the other hand, equation (24.15) tells us that

$$[L_2, L_{-2}] = 4L_0 + \frac{D}{2}\,. \tag{24.23}$$

The left-hand side has been shown to annihilate $|\Phi\rangle$, so the right-hand side must annihilate $|\Phi\rangle$, as well,

$$\Big(4(L_0 - 1) + 4 + \frac{D}{2}\Big)|\Phi\rangle = \Big(4 + \frac{D}{2}\Big)|\Phi\rangle = 0\,. \tag{24.24}$$

Since D is positive, we find $|\Phi\rangle = 0$. This confirms that we cannot impose all the Virasoro conditions and expect to find nontrivial states.

Bearing this result in mind, we attempt to find a subset of the Virasoro constraints that can be imposed without setting to zero the states. The fact that a subset of the original set of constraints may suffice to define a consistent quantum theory is not obvious, but past experience suggests that it may. In the covariant quantization of electromagnetism, for example, the Lorentz gauge condition $\partial \cdot A = 0$ appears as a quantum constraint. This condition cannot be fully imposed on photon states. Instead, only "half" of the gauge condition needs to be imposed on states in order for the theory to work properly. We will follow a similar strategy here.

The complete set of operators that might annihilate physical states is given by

$$\{\ldots, L_{-3}, L_{-2}, L_{-1}\,,\ L_0 - 1\,,\ L_1, L_2, L_3, \ldots\}. \tag{24.25}$$

We have already seen that the constraint $L_0 - 1 = 0$ is necessary to fix the mass spectrum correctly. The simplest thing to try is to set to zero either all operators which have positive mode number or all operators which have negative mode number. In other words, we will impose the constraint

$$(L_0 - 1)|\Phi\rangle = 0 \tag{24.26}$$

together with *one* of the following:

$$L_n|\Phi\rangle \stackrel{?}{=} 0, \quad n > 0, \tag{24.27}$$

$$L_{-n}|\Phi\rangle \stackrel{?}{=} 0, \quad n > 0. \tag{24.28}$$

Only one of these choices works, even though they both have something in common. Both choices result in

$$\langle\Phi|L_n|\Phi\rangle = 0, \quad n \neq 0\,. \tag{24.29}$$

For instance, if we choose (24.27), then (24.29) holds for positive n because L_n kills the state on its right. Furthermore, the Hermitian conjugate of (24.27) yields

$$\langle\Phi|L_{-n} = 0\,, \quad \text{for } n > 0\,, \tag{24.30}$$

or, equivalently,

$$\langle\Phi|L_n = 0\,, \quad \text{for } n < 0\,, \tag{24.31}$$

showing that (24.29) holds for negative n as well. A similar argument can be used for the choice (24.28). So both choices result in physical states for which the *expectation values* of all the L_n with $n \neq 0$ are zero. The condition of vanishing expectation values is weaker than the condition that requires that the L_n annihilate the states. Happily, imposing only half of the constraints does provide a satisfactory theory. It has been proven that half of the constraints suffices to remove the states with negative norms.

We still have to decide which set of conditions to impose. When we were working in light-cone gauge we saw that the positively moded oscillators were the ones that functioned as annihilation operators. The negatively moded oscillators functioned as creation operators. It is natural to associate annihilation with the positively moded Virasoro operators. Therefore we declare the following.

All positively moded Virasoro operators must annihilate physical states. (24.32)

States annihilated by all the positively moded Virasoro operators are called Virasoro *primaries*. We have learned that physical states are Virasoro primaries. Being a Virasoro primary is necessary but not sufficient to guarantee that a state is truly physical. Two additional conditions are necessary. The first is already familiar: the state must be annihilated by $L_0 - 1$. We will call the states that are primary and satisfy this first condition *admissible* states:

$$|\Phi\rangle \text{ is admissible} \iff (L_n - \delta_{n,0})|\Phi\rangle = 0\,, \quad n \geq 0\,. \tag{24.33}$$

The second condition, as we shall see below, has to do with a class of states that are called Virasoro descendents. If a state is admissible and satisfies this additional condition it will be called *physical*. This terminology is not standard: our admissible states are usually called physical, and our physical states are usually called *real* physical states! (Here real is synonymous of truly.)

It is interesting to note that for any fixed state all but a finite number of positively moded Virasoro operators automatically annihilate the state without imposing any condition. Any state $|\Phi_0\rangle$ with number eigenvalue $N_0 \geq 0$ automatically satisfies

$$L_n|\Phi_0\rangle = 0\,, \quad \text{for } n > N_0\,. \tag{24.34}$$

To see this, first note that $L_0 = \alpha' p^2 + N$ implies

$$[N, L_n] = [L_0 - \alpha' p^2, L_n] = [L_0, L_n] = -nL_n\,. \tag{24.35}$$

The mode number of the state $L_n|\Phi_0\rangle$ is therefore

$$N\,L_n|\Phi_0\rangle = [N\,, L_n]\,|\Phi_0\rangle + L_n\,N\,|\Phi_0\rangle = (N_0 - n)L_n|\Phi_0\rangle\,. \tag{24.36}$$

If $n > N_0$, then $L_n|\Phi_0\rangle$ is a state of *negative* number eigenvalue. Since N only has non-negative eigenvalues, the state must vanish.

A Virasoro *descendent* of a given primary is a state that can be written as a finite linear combination of products of negatively moded Virasoro operators, acting on the primary state. For example, if $|p\rangle$ denotes a primary state, the state $L_{-1}|p\rangle$ is a descendent of $|p\rangle$, and so is $(L_{-2}L_{-1} + L_{-4}L_{-3})|p\rangle$. Since descendents play an important role, let us discuss them in some detail.

It follows from (24.36) that $NL_{-1}|p\rangle = (N_p + 1)L_{-1}|p\rangle$, where N_p is the number eigenvalue of the primary $|p\rangle$. The state $L_{-1}|p\rangle$ is, up to scale, the unique descendent of $|p\rangle$ with number $N_p + 1$. There are two basis descendents with number $N_p + 2$: $L_{-2}|p\rangle$ and $L_{-1}L_{-1}|p\rangle$. For descendents with number $N_p + 3$ the counting gets a bit more interesting. A list of candidate basis descendents is

$$L_{-3}|p\rangle\,, \quad L_{-2}L_{-1}|p\rangle\,, \quad L_{-1}L_{-2}|p\rangle\,, \quad (L_{-1})^3|p\rangle\,.$$

Because the Virasoro operators do not commute, the second and third states are not identical. There is, however, one linear relation among the above states, since

$$L_{-1}L_{-2} = [L_{-1}, L_{-2}] + L_{-2}L_{-1} = L_{-3} + L_{-2}L_{-1}\,.$$

This identity allows us to rewrite the third state in terms of the first two, so there are only three basis descendents at this level:

$$\text{descendents with number } N_p + 3\text{:} \quad L_{-3}|p\rangle\,, \quad L_{-2}L_{-1}|p\rangle\,, \quad (L_{-1})^3|p\rangle\,. \tag{24.37}$$

In general, for any fixed number $N_p + n$, one can choose a basis set of descendents such that each basis element is of the form:

$$L_{-n_1}L_{-n_2}\ldots L_{-n_k}|p\rangle\,, \quad \text{where} \quad n_1 \geq n_2 \geq \cdots \geq n_k \quad \text{and} \quad \sum_{i=1}^{k} n_i = n\,. \tag{24.38}$$

This is a useful conventional ordering of the Virasoro operators.

Quick calculation 24.2 Convince yourself that any descendent of $|p\rangle$ with number $N_p + n$ that is written as an arbitrary sequence of negatively moded Virasoro operators acting on $|p\rangle$ can be written as a linear superposition of states of the form (24.38). Note that the number of elements in the generating set (24.38) is equal to the number of partitions of n.

For any given primary $|p\rangle$, there may be linear relations between the basis descendents (24.38). These relations do not arise by manipulation of the Virasoro operators but rather from the specific properties of the state $|p\rangle$. There are simple examples of this phenomenon. The zero-momentum ground state $|0\rangle$, for example, is a primary whose descendent $L_{-1}|0\rangle$ vanishes identically (Problem 24.2). An important property of descendents is that they are all orthogonal to any primary. Indeed, any basis descendent $|d\rangle$ can be written as $|d\rangle = L_{-n_i}|\chi\rangle$ for some $n_i > 0$ and some state $|\chi\rangle$. It follows that for any primary $|p\rangle$,

$$\langle d \,|\, p\rangle = \langle \chi | L_{n_i} | p\rangle = 0\,, \tag{24.39}$$

since $|p\rangle$ is annihilated by all positively moded Virasoro operators.

We now suggest that a state that is both a primary and a descendent represents a state that is pure gauge. A state that is both a primary and a descendent is called a *null* state. It follows from (24.39) that a null state has zero inner product with itself, with any primary, and with any descendent. If we alter a primary state by the addition of a null state, then the new primary state has the same inner products with primary states as the original one. A null state behaves as a pure gauge if, in addition, physical operators in the theory map null states to null states; in this case the addition of null states to primary states cannot affect any physical expectation values. This motivates the following definition of a physical state.

> A nonvanishing state is said to represent a physical state if it is admissible and it is not a descendent, i.e., the state must be a primary, it must be annihilated by $(L_0 - 1)$, and it must not be a descendent. Two representatives of the same physical state must differ by a null state.

Note that we speak about representatives of physical states precisely because of the ambiguity that null states create. A physical state is *not* best thought of as one specific vector in the state space but rather as a class of vectors, all of which differ from one another by a null state. Any vector in this class is an equally valid representative of the *same* physical state. In order to give evidence that this definition is a good one, we will apply it in the following section to find the physical states for the two lowest-mass levels of the open string. We will recover the results previously derived in the light-cone gauge.

24.4 Lorentz covariant state space

Before constructing the state space of the covariantly quantized string, it is useful to recall some properties of the light-cone state space. The light-cone time-independent states are labeled by the light-cone momenta p^+ and $\vec{p}_T$. The ground states, for example, are written

as $|p^+, \vec{p}_T\rangle$. Time-dependent states satisfy the Schrödinger equation which uses the light-cone Hamiltonian, as we saw in Section 11.4 and Section 12.7.

In the covariant formalism all components p^μ of the momentum are independent commuting operators, and we can label the ground states using the full momentum vector $|p\rangle = |p^0, p^1, \ldots, p^{25}\rangle$. The conjugate variables x^μ label the position states $|x\rangle = |x^0, x^1, \ldots, x^{25}\rangle$. There are as many position states as there are *spacetime* points, and the states carry time labels! Given a state $|\psi\rangle$, the associated wavefunction $\langle x|\psi\rangle$ will have time dependence even before we introduce a Schrödinger equation.

To understand the fate of the Schrödinger equation let us examine the Hamiltonian (24.6). Rewriting it in terms of $\dot{X}$, we have

$$\begin{aligned} H &= \frac{1}{4\pi\alpha'}\int_0^\pi d\sigma \left(\dot{X}^2 + X'^2\right) \\ &= \frac{1}{4\pi\alpha'}\int_0^\pi d\sigma \,\frac{1}{2}\left(\left(\dot{X}+X'\right)^2 + \left(\dot{X}-X'\right)^2\right) \\ &= \frac{1}{2\pi}\sum_{n\in\mathbb{Z}}\int_0^\pi d\sigma\, L_n(e^{-in\sigma} + e^{in\sigma})\, e^{-in\tau}\,, \end{aligned} \tag{24.40}$$

where we used (24.13). All integrals with $n \neq 0$ vanish, and for $n = 0$ the right-hand side gives L_0. This operator, however, is ambiguous because (24.13) does not yield the normal-ordered L_0 operator. Thus H equals L_0 up to a constant. The light-cone Hamiltonian is $L_0^\perp - 1$, suggesting that the covariant Hamiltonian must be chosen to be

$$H = L_0 - 1 = \alpha' p^2 + N - 1\,. \tag{24.41}$$

This ordering constant coincides with the one needed for the constraint associated with L_0. Since all physical states are annihilated by $L_0 - 1$, we reach the surprising conclusion that the Hamiltonian annihilates all physical states! This means that it is not possible to introduce a time variable and generate nontrivial time evolution through the Schrödinger equation. But then again, we do not need to introduce a time variable and generate time evolution as we did in the light-cone. The covariant states already have time labels. The Schrödinger equation has turned into the constraint $H|\Phi\rangle = 0$. This constraint – roughly speaking – fixes a relation between the time label and the position labels of the states. In momentum space, it imposes a mass-shell condition that fixes the energy p^0 in terms of the momentum $\vec{p}$.

The basis vectors of the covariant state space are constructed by acting on the ground states $|p\rangle$ with all possible creation operators:

$$|r\rangle = \prod_{n=1}^{\infty}\prod_{\mu=0}^{25}\left(a_n^{\mu\,\dagger}\right)^{\lambda_{n,\mu}}|p\rangle\,. \tag{24.42}$$

Here the $\lambda_{n,\mu}$ are non-negative integers, and, as usual, only finitely many of them are non-zero. All 26 values of μ can be used to build the basis states. In the light-cone gauge all the basis states we introduced were physical states. Not here: a vector $|\Phi\rangle$ represents a physical state if it satisfies the Virasoro constraints in (24.33) and it is not a descendent.

Let us examine concretely some of the physical states in the theory. Consider first the ground states $|p\rangle$. These are the states with number eigenvalue zero. On account of (24.34), they are automatically annihilated by the $L_{n\geq 1}$ operators, and the only nontrivial constraint is

$$0 = (L_0 - 1)|p\rangle = (\alpha' p^2 - 1)|p\rangle \;\longrightarrow\; p^2 = 1/\alpha' \,. \tag{24.43}$$

This is the on-shell condition for a tachyon state with $M^2 = -1/\alpha'$. The states $|p\rangle$ are not descendents because if they were, they would have to have descended from a state with negative number eigenvalue, and there are no such states in the theory. We have therefore identified the physical tachyon states. In Chapter 12 we identified the same physical states using the light-cone quantization of the open string.

Quick calculation 24.3 Use (24.13) to verify explicitly that $L_1|p\rangle = 0$.

With a slight generalization we can also deal with tachyon *fields*, and the constraint will give us the classical field equation for the tachyon. A general tachyon state $|T\rangle$ can be constructed as a superposition of momentum ground states:

$$|T\rangle \equiv \int d^D p\, \phi(p)\, |p\rangle \,, \tag{24.44}$$

where $\phi(p)$ is an arbitrary function of the spacetime momentum. For $|T\rangle$ to be physical it must be annihilated by $(L_0 - 1)$:

$$(L_0 - 1)|T\rangle = \int d^D p\, \phi(p)(L_0 - 1)|p\rangle = \int d^D p\, \phi(p) \left(\alpha' p^2 - 1\right) |p\rangle = 0 \,. \tag{24.45}$$

Since the ground states $|p\rangle$ are linearly independent, the integrand must vanish for all p:

$$(\alpha'\, p^2 - 1)\phi(p) = 0 \,. \tag{24.46}$$

This is the field equation for the tachyon. The function $\phi(p)$, which was introduced to help construct the general tachyon state, turns out to be the tachyon field. Since the tachyon state $|T\rangle$ has number eigenvalue zero, there are no further Virasoro constraints, and the state cannot be a descendent.

Next we consider photon states, which we identified as $\xi_I\, a_1^{I\dagger}|p^+, \vec{p}_T\rangle$ in the light-cone quantization of the open string. Presently, we consider $N = 1$ states with fixed momentum p and polarization ξ_μ:

$$\xi_\mu\, \alpha_{-1}^\mu |p\rangle \,. \tag{24.47}$$

The $L_0 - 1 = 0$ condition gives

$$0 = (\alpha' p^2 + N - 1)\, \xi_\mu\, \alpha_{-1}^\mu |p\rangle = \alpha' p^2 \xi_\mu\, \alpha_{-1}^\mu |p\rangle \;\longrightarrow\; p^2 = 0 \,. \tag{24.48}$$

With $L_1 = \alpha_0 \cdot \alpha_1 + (\alpha_{-1} \cdot \alpha_2 + \alpha_{-2} \cdot \alpha_3 + \cdots)$, the condition $L_1 \xi_\mu\, \alpha_{-1}^\mu |p\rangle = 0$ gives

$$0 = \alpha_0 \cdot \alpha_1\, \xi_\mu\, \alpha_{-1}^\mu |p\rangle = \sqrt{2\alpha'}\; p^\mu\, \xi_\mu\, |p\rangle \;\longrightarrow\; p \cdot \xi = 0 \,. \tag{24.49}$$

The $L_{n\geq 2}$ operators automatically annihilate the $\xi_\mu\, \alpha_{-1}^\mu |p\rangle$ states. So far, we have learned that physical photon states exist only for $p^2 = 0$ and must satisfy $p \cdot \xi = 0$. For any p_μ

which satisfies $p^2 = 0$, we can choose a Lorentz frame where $p_\mu = (p_0, p_0, 0, \ldots, 0)$, and then the constraint $p \cdot \xi = 0$ gives

$$\xi^0 + \xi^1 = 0\,. \tag{24.50}$$

This cannot be the final answer; (24.50) leaves $D-1$ independent polarizations ξ_μ, and our light-cone analysis showed that photons only have $D-2$ independent polarizations. We must impose the condition that physical states are defined up to null states. For this, consider the descendent

$$|d\rangle = L_{-1}\frac{1}{\sqrt{2\alpha'}}\, i\epsilon\,|p\rangle = \big(ip_\mu\epsilon\big)\,\alpha^\mu_{-1}|p\rangle\,, \quad \text{with} \quad p^2 = 0. \tag{24.51}$$

This state is of the form (24.47) with $\xi_\mu = ip_\mu\epsilon$. It follows from (24.49) that $|d\rangle$ is also primary since $p^\mu(i\epsilon p_\mu) = 0$. So $|d\rangle$ is null. We therefore have the equivalence of states

$$\xi_\mu\,\alpha^\mu_{-1}\,|p\rangle \;\sim\; \big(\xi_\mu + ip_\mu\epsilon\big)\,\alpha^\mu_{-1}\,|p\rangle, \tag{24.52}$$

and, consequently, the equivalence of polarizations

$$\xi^\mu \sim \xi^\mu + ip^\mu\epsilon\,, \quad \text{for} \quad p^2 = 0\,. \tag{24.53}$$

Again, using $p_\mu = (p_0, p_0, 0, \ldots, 0)$, this means that

$$\xi^0 \sim \xi^0 - ip_0\,\epsilon \quad \text{and} \quad \xi^1 \sim \xi^1 + ip_0\,\epsilon\,. \tag{24.54}$$

Since $\xi^0 - \xi^1 \sim \xi^0 - \xi^1 - 2ip_0\epsilon$, we can choose representative polarizations such that

$$\xi^0 - \xi^1 = 0\,. \tag{24.55}$$

Together with (24.50), we conclude that physical states have $D-2$ independent polarizations. This is the result we had found in the light-cone gauge.

One can also work with general superpositions. We define a gauge field state $|A\rangle$ as

$$|A\rangle = \int d^D p\, A_\mu(p)\,\alpha^\mu_{-1}\,|p\rangle\,. \tag{24.56}$$

Quick calculation 24.4 Show that the $L_0 - 1$ and L_1 constraints give $p^2 A_\mu(p) = 0$ and $p \cdot A = 0$, respectively. The condition $p \cdot A = 0$ is the Lorentz gauge condition. The equation $p^2 A_\mu = 0$ is the familiar Maxwell field equation $p^2 A^\mu - p^\mu(p \cdot A) = 0$ in the Lorentz gauge.

For additional discussion of photon states, see Problem 24.3.

24.5 Closed string Virasoro operators

The covariant quantization of the closed string brings about no new complications. There are two sets of covariant Virasoro operators, and the operators with non-negative mode number annihilate the physical states. The vanishing of both Virasoro operators with mode

number zero (with a proper subtraction constant) implies that the left and right number operators have the same eigenvalue when acting on physical states. We will only consider closed strings in the absence of compactification.

The mode expansion (13.24) for a generic closed string coordinate reads

$$X^\mu(\tau,\sigma) = x_0^\mu + \sqrt{2\alpha'}\,\alpha_0^\mu \tau + i\sqrt{\frac{\alpha'}{2}}\sum_{n\neq 0}\frac{e^{-in\tau}}{n}(\alpha_n^\mu e^{in\sigma} + \bar\alpha_n^\mu e^{-in\sigma})\,, \tag{24.57}$$

where $\alpha_0^\mu = p^\mu\sqrt{\alpha'/2}$. The τ and σ derivatives of the coordinates were recorded in (13.26), with the understanding that $\bar\alpha_0^\mu = \alpha_0^\mu$. Just as in (13.36), we have

$$\begin{aligned}(\dot X^\mu + X^{\mu\prime})^2 &= 4\alpha'\sum_{n\in\mathbb{Z}}\Big(\frac{1}{2}\sum_{p\in\mathbb{Z}}\bar\alpha_p^\mu\bar\alpha_{n-p,\mu}\Big)e^{-in(\tau+\sigma)} \equiv 4\alpha'\sum_{n\in\mathbb{Z}}\bar L_n e^{-in(\tau+\sigma)}\,,\\ (\dot X^\mu - X^{\mu\prime})^2 &= 4\alpha'\sum_{n\in\mathbb{Z}}\Big(\frac{1}{2}\sum_{p\in\mathbb{Z}}\alpha_p^\mu\alpha_{n-p,\mu}\Big)e^{-in(\tau-\sigma)} \equiv 4\alpha'\sum_{n\in\mathbb{Z}} L_n e^{-in(\tau-\sigma)}\,,\end{aligned} \tag{24.58}$$

where we defined

$$\bar L_n = \frac{1}{2}\sum_{p\in\mathbb{Z}}\bar\alpha_p^\mu\bar\alpha_{n-p,\mu}\,,\qquad L_n = \frac{1}{2}\sum_{p\in\mathbb{Z}}\alpha_p^\mu\alpha_{n-p,\mu}\,. \tag{24.59}$$

These are the two sets of Virasoro operators of closed string theory. If we imposed the constraints (24.1), then all Virasoro operators would have to annihilate the physical states. Just like for open strings, this would leave no physical states. In analogy with the open string quantum constraints (24.33), we demand that for physical closed states $|\Psi\rangle$

$$(L_n - \delta_{n,0})|\Psi\rangle = 0\,,\quad (\bar L_n - \delta_{n,0})|\Psi\rangle = 0\,,\quad n\geq 0\,. \tag{24.60}$$

A vector $|\Psi\rangle$ which satisfies these conditions is a representative for a physical closed string state if it is not a descendent. A descendent of a primary $|p\rangle$ is a state obtained by acting on the primary with a collection of negatively moded Virasoro operators. Both barred and unbarred operators can be used.

For $n = 0$, the conditions (24.60) give

$$\begin{aligned}(L_0 - 1)|\Psi\rangle &= \Big(\frac{\alpha'}{4}p^2 + N - 1\Big)|\Psi\rangle = 0\,,\\ (\bar L_0 - 1)|\Psi\rangle &= \Big(\frac{\alpha'}{4}p^2 + \bar N - 1\Big)|\Psi\rangle = 0\,,\end{aligned} \tag{24.61}$$

where N and $\bar N$ are the covariant number operators. Note that equations (24.61) imply that $L_0 - \bar L_0$ annihilates all physical states. As before, this condition is interpreted as an invariance of physical states under a constant shift of the σ coordinate on the world-sheet,

or as the vanishing of the two-dimensional momentum along the string (Section 13.2). Equations (24.61) give *two* expressions for the mass-squared:

$$M^2 = -p^2 = \frac{4}{\alpha'}(N-1) = \frac{4}{\alpha'}(\bar{N}-1)\,, \tag{24.62}$$

and therefore we have the constraint

$$N = \bar{N}\,. \tag{24.63}$$

Making use of (24.63) we can write, more symmetrically,

$$M^2 = \frac{2}{\alpha'}(N+\bar{N}-2)\,. \tag{24.64}$$

The closed string state space is constructed by acting with arbitrary numbers of oscillators on the closed string ground states $|p\rangle$. To uncover the closed string tachyon states it suffices to consider the ground states and to impose the $L_0 - 1 = \bar{L}_0 - 1 = 0$ conditions. To uncover the massless closed string states – the Kalb–Ramond field, the graviton, and the dilaton – we must also examine the $L_1 = \bar{L}_1 = 0$ constraints. Those constraints suffice because the massless states have $N = \bar{N} = 1$, so Virasoro operators with higher mode number automatically annihilate these states. Of course, the equivalences generated by null states must also be considered (Problem 24.5).

24.6 The Polyakov string action

Despite the elegance of the action (24.4) and the ease with which we can obtain from it the equations of motion, the momenta, and the Hamiltonian, the constraint equations (24.1) still have to be imposed by hand. We will now develop another action, the Polyakov action, from which the constraint equations also emerge naturally. As a first step, we rewrite (24.4) in a more suggestive form:

$$S = -\frac{1}{4\pi\alpha'}\int d\tau d\sigma\, \eta^{\alpha\beta}\partial_\alpha X^\mu \partial_\beta X^\nu\, \eta_{\mu\nu}\,. \tag{24.65}$$

The spacetime indices on the string coordinates are contracted with the Minkowski metric $\eta_{\mu\nu}$. The α and β indices run over two values, corresponding to the two world-sheet coordinates τ and σ:

$$\partial_\alpha = \frac{\partial}{\partial\xi^\alpha}\,, \quad \xi^\alpha = (\xi^1, \xi^2) = (\tau, \sigma). \tag{24.66}$$

We have also introduced a *two-dimensional* Minkowski metric $\eta^{\alpha\beta}$. Just like the spacetime Minkowski metric, it is diagonal; the time–time entry is -1, and the space–space entry is $+1$:

$$\eta^{\alpha\beta} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{24.67}$$

In (24.65) the repeated α and β indices are summed over the two values τ and σ. Writing out these sums we recover the action (24.4).

The above two-dimensional metric might remind you of the metric that emerged while we were making manifest the reparameterization invariance of the Nambu–Goto action. There we saw that the action takes the form indicated in equation (6.44):

$$S = -\frac{1}{2\pi\alpha'}\int d\tau d\sigma \sqrt{-\gamma}\,. \tag{24.68}$$

Here $\gamma = \det(\gamma_{\alpha\beta})$, and $\gamma_{\alpha\beta}$ is the world-sheet metric induced by the target-space Minkowski metric. It is explicitly given by (6.42):

$$\gamma_{\alpha\beta} = \frac{\partial X}{\partial \xi^\alpha}\cdot\frac{\partial X}{\partial \xi^\beta} = \partial_\alpha X\cdot\partial_\beta X\,. \tag{24.69}$$

The Polyakov action involves a new world-sheet metric, $h_{\alpha\beta}(\tau,\sigma)$. This is the kind of metric that one uses in two-dimensional general relativity. The metric $h_{\alpha\beta}(\tau,\sigma)$ is a dynamical variable in the action, so it leads to its own field equations. As it turns out, the equations of motion will relate the new metric $h_{\alpha\beta}$ to the induced metric $\gamma_{\alpha\beta}$. The metric $h_{\alpha\beta}$ enters the Polyakov action in a way that is analogous to the way in which $\eta^{\alpha\beta}$ enters the action (24.65). The Polyakov action is

$$S = -\frac{1}{4\pi\alpha'}\int d\tau d\sigma\,\sqrt{-h}\,h^{\alpha\beta}\partial_\alpha X^\mu\partial_\beta X^\nu\eta_{\mu\nu}\,. \tag{24.70}$$

Here $h \equiv \det(h_{\alpha\beta})$, and $h^{\alpha\beta}$ is the inverse of $h_{\alpha\beta}$:

$$h^{\alpha\beta}h_{\beta\gamma} = \delta^\alpha_\gamma\,,\quad h^{\alpha\beta}h_{\alpha\beta} = 2\,. \tag{24.71}$$

The variation of the metric $h_{\alpha\beta}$ in the Polyakov action will give us the Virasoro constraints. Since $h_{\alpha\beta}$ is a symmetric two-by-two matrix, and such matrices have three independent entries, we should expect three constraints. But there are only two Virasoro conditions. How can these facts be reconciled?

First note that the metric $h^{\alpha\beta}$ enters the action through the specific combination $\sqrt{-h}\,h^{\alpha\beta}$. The square root factor is needed to make the measure $(d\tau d\sigma\sqrt{-h}\,)$ reparameterization invariant, as you saw in Section 6.2. The $h^{\alpha\beta}$ factor is needed to contract against the indices carried by the derivatives of X. While $h^{\alpha\beta}$ is clearly determined by three real numbers at any fixed point on the world-sheet, the combination $\sqrt{-h}\,h^{\alpha\beta}$ is in fact determined by just two numbers! This is a peculiar property of two-dimensional metrics. Define this combination to be

$$M^{\alpha\beta} = \sqrt{-h}\,h^{\alpha\beta}. \tag{24.72}$$

We will show that the symmetric matrix M satisfies an additional constraint. To appreciate why two-dimensional metrics are special, let n denote the size of the matrix $h^{\alpha\beta}$. Then

$$\det(M^{\alpha\beta}) = (\sqrt{-h}\,)^n\det(h^{\alpha\beta}) = \frac{(-h)^{\frac{n}{2}}}{\det(h_{\alpha\beta})} = \frac{(-h)^{\frac{n}{2}}}{h} = -(-h)^{\frac{n}{2}-1}\,. \tag{24.73}$$

The final simplification occurs only when $n = 2$. In that case we find

$$\det(M^{\alpha\beta}) = -1\,. \tag{24.74}$$

A 2-by-2 symmetric matrix with determinant equal to a fixed constant is determined by two parameters. This is why the variation of the metric gives us only two independent constraints, which is the number we want.

Enough with preliminaries. Let us tackle the variation of the Polyakov action. First we vary the string coordinates X^μ. This variation gives

$$\begin{aligned}\delta S &= -\frac{1}{2\pi\alpha'}\int d\tau d\sigma \sqrt{-h}\, h^{\alpha\beta}\partial_\alpha(\delta X^\mu)\partial_\beta X^\nu \eta_{\mu\nu} \\ &= \frac{1}{2\pi\alpha'}\int d\tau d\sigma\ \delta X^\mu\ \partial_\alpha\Big(\sqrt{-h}\, h^{\alpha\beta}\partial_\beta X^\nu \eta_{\mu\nu}\Big),\end{aligned} \tag{24.75}$$

where we have discarded the complete derivatives which give the familiar boundary contributions in the case of open strings. The resulting equation of motion is therefore

$$\partial_\alpha\Big(\sqrt{-h}\, h^{\alpha\beta}\partial_\beta X^\mu\Big) = 0. \tag{24.76}$$

This equation is not yet a wave equation for X^μ, but we will get there shortly.

In order to vary the metric $h^{\alpha\beta}$ we need a preliminary result. Let A be a 2-by-2 matrix, and let δA denote its variation:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22}\end{pmatrix} \quad \text{and} \quad \delta A = \begin{pmatrix} \delta a_{11} & \delta a_{12} \\ \delta a_{21} & \delta a_{22}\end{pmatrix}. \tag{24.77}$$

It is straightforward to check that the variation of $\det A$ can be written as

$$\delta \det A = (\det A)\,\mathrm{Tr}(A^{-1}\,\delta A)\,, \tag{24.78}$$

where A^{-1} denotes the inverse of A and Tr stands for trace. This identity actually holds for matrices of arbitrary dimension.

Quick calculation 24.5 Verify explicitly that equation (24.78) holds for 2-by-2 matrices.

Let $\delta h^{\alpha\beta}$ denote the variation of the metric. We can use (24.78) to calculate the variation of h:

$$\delta h = \delta \det(h_{\alpha\beta}) = h\,(h^{\alpha\beta}\delta h_{\beta\alpha})\,. \tag{24.79}$$

We must write the variation $\delta h_{\alpha\beta}$ in terms of $\delta h^{\alpha\beta}$. For this we vary the second equation in (24.71) to find

$$\delta h^{\alpha\beta}\ h_{\alpha\beta} + h^{\alpha\beta}\,\delta h_{\alpha\beta} = 0 \ \longrightarrow\ h^{\alpha\beta}\,\delta h_{\beta\alpha} = -\delta h^{\alpha\beta}\ h_{\alpha\beta}\,, \tag{24.80}$$

where we noted that $\delta h_{\alpha\beta} = \delta h_{\beta\alpha}$. We can then rewrite the variation of h as

$$\delta h = -h\,\delta h^{\alpha\beta}\,h_{\alpha\beta}\,. \tag{24.81}$$

Of course, what we really need to vary is $\sqrt{-h}$. This can now be done as

$$\delta(\sqrt{-h}) = -\frac{1}{2}\frac{\delta h}{\sqrt{-h}} = -\frac{1}{2}\frac{(-h)\delta h^{\alpha\beta}h_{\alpha\beta}}{\sqrt{-h}} = -\frac{1}{2}\sqrt{-h}\ \delta h^{\alpha\beta}\,h_{\alpha\beta}\,. \tag{24.82}$$

We can finally vary the metric in the action (24.70):

$$\begin{aligned}\delta S &= -\frac{1}{4\pi\alpha'}\int d\tau d\sigma\sqrt{-h}\Big(-\frac{1}{2}\delta h^{\alpha\beta}\,h_{\alpha\beta}(h^{\gamma\delta}\partial_\gamma X\cdot\partial_\delta X)+\delta h^{\alpha\beta}\partial_\alpha X\cdot\partial_\beta X\Big)\\ &= -\frac{1}{4\pi\alpha'}\int d\tau d\sigma\sqrt{-h}\,\delta h^{\alpha\beta}\Big(\partial_\alpha X\cdot\partial_\beta X-\frac{1}{2}h_{\alpha\beta}(h^{\gamma\delta}\partial_\gamma X\cdot\partial_\delta X)\Big)\,. \end{aligned}\tag{24.83}$$

The equation of motion which arises from the variation $\delta h^{\alpha\beta}$ is therefore

$$\partial_\alpha X\cdot\partial_\beta X-\frac{1}{2}h_{\alpha\beta}(h^{\gamma\delta}\partial_\gamma X\cdot\partial_\delta X)=0\,.\tag{24.84}$$

Recalling that $\partial_\alpha X\cdot\partial_\beta X$ is just the induced metric $\gamma_{\alpha\beta}$ in (24.69), the above equation can be written as

$$\gamma_{\alpha\beta}-\frac{1}{2}h_{\alpha\beta}(h^{\gamma\delta}\gamma_{\gamma\delta})=0\,.\tag{24.85}$$

Since the factor $(h^{\gamma\delta}\gamma_{\gamma\delta})$ has no free indices, this equation sets the world-sheet metric $h_{\alpha\beta}$ proportional to the induced metric $\gamma_{\alpha\beta}$ at every point on the world-sheet. The factor of proportionality can be position dependent. This factor is *not* determined by (24.85): given a solution $h_{\alpha\beta}$ of this equation, the metric $h'_{\alpha\beta}=\Omega^2 h_{\alpha\beta}$, with Ω an arbitrary nonvanishing function, will also provide a solution. This happens because the second term on the left-hand side of (24.85) contains the product of the metric and its inverse. Equation (24.85) is therefore satisfied by

$$h_{\alpha\beta}=f^2(\xi)\,\gamma_{\alpha\beta}\,,\tag{24.86}$$

where $f(\xi)$ is some undetermined nonvanishing function on the world-sheet. By writing (24.86) with $f^2(\xi)$, we are making an additional statement: the proportionality factor that relates the metrics is positive. This has a physical implication: the notions of timelike and spacelike vectors defined by $h_{\alpha\beta}$ and $\gamma_{\alpha\beta}$ agree. Since the induced metric $\gamma_{\alpha\beta}$ is really the ambient metric referred to the world-sheet, if a world-sheet vector is spacelike or timelike, as determined by $h_{\alpha\beta}$, it will be spacelike or timelike, respectively, as determined by the Minkowski metric. Metrics related by a (positive) factor of proportionality are said to be conformal to each other. Thus the world-sheet metric is conformal to the induced metric.

Let us see what happens to the Polyakov action when we substitute into it the information we have gained. We first calculate $\sqrt{-h}h^{\alpha\beta}=M^{\alpha\beta}$ using (24.86). The good news is that the undetermined function $f(\xi)$ drops out of this calculation. To see this quickly, recall that the determinant of the matrix $M^{\alpha\beta}$ is (-1), independent of the metric $h^{\alpha\beta}$. If we scale the metric $h^{\alpha\beta}$, the most that could happen to $M^{\alpha\beta}$ is that it scales as well. But if $M^{\alpha\beta}$ were to scale then its determinant would change. Thus the scale of $h^{\alpha\beta}$ does not affect $M^{\alpha\beta}$. This argument can be confirmed explicitly. With $\gamma\equiv\det(\gamma_{\alpha\beta})$, we use (24.86) to determine the determinant of $h_{\alpha\beta}$ and the inverse metric $h^{\alpha\beta}$:

$$h=f^4\,\gamma\quad\text{and}\quad h^{\alpha\beta}=\frac{1}{f^2}\,\gamma^{\alpha\beta}\,,\tag{24.87}$$

where $\gamma^{\alpha\beta}$ is the inverse of $\gamma_{\alpha\beta}$. It follows that

$$\sqrt{-h}\, h^{\alpha\beta} = f^2 \sqrt{-\gamma}\ \frac{1}{f^2}\, \gamma^{\alpha\beta} = \sqrt{-\gamma}\, \gamma^{\alpha\beta}\,, \tag{24.88}$$

which confirms that the undetermined function f is irrelevant. If we now substitute back into the Polyakov action (24.70) we find

$$S = -\frac{1}{4\pi\alpha'} \int d\tau d\sigma \sqrt{-\gamma}\, \gamma^{\alpha\beta}\, \gamma_{\alpha\beta}\,. \tag{24.89}$$

The second equation in (24.71), which is valid for any two-dimensional metric, now gives

$$S = -\frac{1}{2\pi\alpha'} \int d\tau d\sigma \sqrt{-\gamma}\,. \tag{24.90}$$

This is exactly the Nambu–Goto action. We conclude that the Polyakov action is classically equivalent to the Nambu–Goto action.

Quick calculation 24.6 Verify that, had we chosen $h_{\alpha\beta} = -f^2(\xi)\, \gamma_{\alpha\beta}$ instead of (24.86), the Nambu–Goto action would have emerged with the wrong sign.

While we were showing that the factor $f^2(\xi)$ in (24.86) drops out during the evaluation of the string action, we also proved that the Polyakov action is invariant under the transformation

$$h_{\alpha\beta}(\tau, \sigma) \ \rightarrow\ \Omega^2(\tau, \sigma)\, h_{\alpha\beta}(\tau, \sigma)\,, \tag{24.91}$$

where Ω^2 is an arbitrary function on the world-sheet. This rescaling of the world-sheet metric is called a *Weyl transformation*. The invariance of the Polyakov action under a Weyl transformation indicates that, on the world-sheet, distances measured with the $h_{\alpha\beta}$ metric have no physical significance.

We have yet to show that we obtain the Virasoro constraints and wave equations for the string coordinates. The equation of motion (24.76) for X^μ looks complicated, and the equation of motion (24.84) for $h_{\alpha\beta}$ does not resemble the Virasoro constraints. In order to make progress we use the reparameterization invariance to choose a convenient form for $h_{\alpha\beta}$.

It is a well known result of two-dimensional geometry that a coordinate reparameterization allows an arbitrary metric $h_{\alpha\beta}$ on a surface to be cast locally in the form

$$h_{\alpha\beta} = \rho^2(\xi)\, \eta_{\alpha\beta}\,. \tag{24.92}$$

Here ρ is a *conformal factor* and $\eta_{\alpha\beta}$ is the two-dimensional Minkowski metric. The metric $h_{\alpha\beta}$ is said to be conformally flat. By restricting ourselves to the class of world-sheet coordinates that result in a conformally flat metric, we are making a partial gauge choice. This choice is called the conformal gauge. In the conformal gauge $\sqrt{-h}h^{\alpha\beta} = \eta^{\alpha\beta}$, so equation (24.76) becomes

$$\partial_\alpha\left(\eta^{\alpha\beta}\partial_\beta X^\mu\right) = \eta^{\alpha\beta}\partial_\alpha\partial_\beta X^\mu = 0\,. \tag{24.93}$$

Expanding out the sums over two-dimensional indices, we see that this is precisely the wave equation. For equations (24.84), the conformal gauge condition (24.92) results in

$$\partial_\alpha X \cdot \partial_\beta X - \frac{1}{2}\,\eta_{\alpha\beta}\,(\eta^{\gamma\delta}\partial_\gamma X \cdot \partial_\delta X) = 0\,. \tag{24.94}$$

Expanding the expression in parentheses we have

$$\partial_\alpha X \cdot \partial_\beta X - \frac{1}{2}\,\eta_{\alpha\beta}\,(-\dot{X}^2 + X'^2) = 0\,. \tag{24.95}$$

These are three equations, but only two of them are independent. Taking $\alpha = \beta = 1$ we get

$$\dot{X}^2 + \frac{1}{2}(-\dot{X}^2 + X'^2) = 0 \;\longrightarrow\; \dot{X}^2 + X'^2 = 0\,, \tag{24.96}$$

which is one of the constraints. With $\alpha = 1, \beta = 2$ we get

$$\dot{X} \cdot X' = 0\,, \tag{24.97}$$

which is the second constraint. Finally, when $\alpha = \beta = 2$ we get

$$X'^2 - \frac{1}{2}(-\dot{X}^2 + X'^2) = 0 \;\longrightarrow\; \dot{X}^2 + X'^2 = 0\,, \tag{24.98}$$

which is redundant. We have thus shown that the equations of motion for the world-sheet metric reduce to the familiar Virasoro constraints in the conformal gauge.

Because of its elegance and convenience, the Polyakov string action is typically the starting point of any detailed analysis of covariant string quantization. In such an analysis, the various facts for which we have had to appeal to our previous light-cone results can be given independent justification.

Problems

Problem 24.1 Covariant quantization of the point particle.

The point particle Lagrangian is $L = -m\sqrt{-\dot{x}^2}$, where $\dot{x}^\mu = dx^\mu(\tau)/d\tau$.

(a) Calculate the momentum p_μ canonical to $x^\mu(\tau)$. Verify that the momentum satisfies the constraint $p^2 + m^2 = 0$. Prove that one cannot solve uniquely for the velocity in terms of the momentum by showing that, given one solution, one can easily construct a different one. In addition, only for momenta satisfying the constraint can one possibly find velocities.

(b) Show that $H \equiv p_\mu \dot{x}^\mu - L = 0$. Set up commutation relations and describe the state space.

(c) Physical states are those that satisfy the constraint. Build candidate states as general linear superpositions and compare with (11.48). Apply the constraint and show that the wavefunction for the physical states satisfies the Klein–Gordon equation for a scalar field of mass m.

Problem 24.2 States as Virasoro primaries or Virasoro descendents.

Consider the Virasoro operators $L_n = \frac{1}{2}\sum_p \alpha_{n-p}\alpha_p$ associated with a single open string coordinate X with oscillators that satisfy $[\alpha_m, \alpha_n] = m\delta_{m+n,0}$. The states that can be built from the zero-momentum vacuum $|0\rangle$ for $N \leq 4$ are:

$$\begin{aligned}
N = 0: &\quad |0\rangle\,, \\
N = 1: &\quad \alpha_{-1}|0\rangle\,, \\
N = 2: &\quad \alpha_{-2}|0\rangle\,,\ \ \alpha_{-1}\alpha_{-1}|0\rangle\,, \\
N = 3: &\quad \alpha_{-3}|0\rangle\,,\ \ \alpha_{-2}\alpha_{-1}|0\rangle\,,\ \ (\alpha_{-1})^3|0\rangle\,, \\
N = 4: &\quad \alpha_{-4}|0\rangle\,,\ \ \alpha_{-2}(\alpha_{-1})^2|0\rangle\,,\ \ \alpha_{-3}\alpha_{-1}|0\rangle\,,\ \ \alpha_{-2}\alpha_{-2}|0\rangle\,,\ \ (\alpha_{-1})^4|0\rangle\,. \qquad (1)
\end{aligned}$$

We want to show that at each level N we can form an equivalent basis of states where each state is either a Virasoro primary or a Virasoro descendent from a primary in the list.

(a) Define the norm of $|0\rangle$ as $\langle 0|0\rangle \equiv 1$. Explain why any state $|\chi\rangle$ in (1) has positive norm $\langle\chi|\chi\rangle > 0$, and why any two different states $|\chi_1\rangle$, $|\chi_2\rangle$ are orthogonal: $\langle\chi_2|\chi_1\rangle = 0$. Argue that any nonzero linear combination of the above states is a state of positive norm, and that, as a result, no state in the state space is both a primary and a descendent.

(b) Explain why the state $|0\rangle$ is a primary. Now consider the state $\alpha_{-1}|0\rangle$ with $N = 1$. This state is either a descendent of $|0\rangle$ or a primary. Show that the only candidate descendent $L_{-1}|0\rangle$ vanishes. Prove that $\alpha_{-1}|0\rangle$ is a primary, and call it $|p_1\rangle$. Using a basis as in equation (24.38), explain why descendents of $|0\rangle$ with one or more L_{-1} operators vanish.

(c) Show that, for $N = 2$, one state is a descendent of $|0\rangle$, and one state is a descendent of $|p_1\rangle$.

(d) Show that, for $N = 3$, one state is a descendent of $|0\rangle$, and two states are descendents of $|p_1\rangle$.

(e) For $N = 4$ we have three candidate descendents of $|p_1\rangle$. From these states, show that one linear combination vanishes:

$$\left(L_{-3} - 2L_{-2}L_{-1} + \tfrac{1}{2}(L_{-1})^3\right)|p_1\rangle = 0\,,$$

and the other two give states equivalent to the first two of the $N = 4$ list. Note that there are only two candidate descendents of $|0\rangle$. So the last three states on the list must break into two descendents and one new primary $|p_4\rangle$. Show that, up to an arbitrary normalization,

$$|p_4\rangle = \left(\alpha_{-3}\alpha_{-1} - \tfrac{3}{4}(\alpha_{-2})^2 - \tfrac{1}{2}(\alpha_{-1})^4\right)|0\rangle\,,$$

and write explicitly the three last states on the list as a linear superposition of this primary and the two descendents.

Problem 24.3 Photon states in the covariant description.

(a) Consider the field equation and gauge transformations for a Maxwell field:

$$p^2 A^\mu - p^\mu (p \cdot A) = 0 \,, \quad \delta A^\mu = i p^\mu \epsilon \,.$$

Show that for $p^2 \neq 0$ it is possible to make the potentials satisfy the Lorentz gauge condition $p \cdot A = 0$. Show also that for $p^2 = 0$ the quantity $p \cdot A$ is gauge invariant and that it is set to zero by the field equation. This shows that the Lorentz gauge condition is generally valid for all p. It also shows, however, that for $p^2 = 0$ the Lorentz gauge does not fix the gauge symmetry. We have seen evidence of this: for $p^2 = 0$ the solutions of the equations of motion in the Lorentz gauge are characterized by $(D-1)$ independent degrees of freedom.

(b) Consider the coupling $\int A_\mu j^\mu d^D x$ of the gauge field to the conserved current j^μ. Show that for $p^2 = 0$ the part of the gauge field that can be gauged away drops out of this coupling.

(c) Examine the descendent $|D\rangle$ relevant to the discussion of the general superposition (24.56):

$$|D\rangle = L_{-1} \frac{1}{\sqrt{2\alpha'}} \int d^D p \; i\epsilon(p) |p\rangle \,.$$

When is $|D\rangle$ null? Show that for $p^2 = 0$ we have the identification $A^\mu(p) \sim A^\mu(p) + i p^\mu \epsilon(p)$. This is an on-shell gauge invariance.

Problem 24.4 D0-branes, open strings, and M-theory.

(a) We want to show that it is not possible to have only one string endpoint on a D0-brane. Intuitively, the string charge, visualized as a current on the string, has nowhere to go on the pointlike D0-brane. More quantitatively, as in any D-brane, on the D0-brane there is a gauge field that couples to the string endpoint as in (16.54). Since the D0-brane has no spatial coordinate the gauge field is just A_0. Show that the Maxwell action vanishes, and that, as a result, the variation of A_0 imposes an inconsistent equation of motion. Note that the inconsistency is removed if the two endpoints of an open string lie on the D0-brane.

(b) Consider the covariant quantization of an open string whose endpoints lie on a D0-brane. Describe the ground states, noting that the momentum has a single component. Construct the general states with $N = 0$ and discuss the equation of motion for the tachyon field $\phi(t)$. Construct the $N = 1$ physical states. Show that there are no relevant null states, and that we have D independent physical states of zero momentum.

(c) Give the mass m_0 of a D0-brane (see Section 18.4). In type IIA superstring theory the same formula applies, and a bound state of n D0-branes is known to have a mass exactly equal to $n m_0$. Such states can be identified with momentum states that arise from a compact eleventh dimension, assuming that the momentum p along this dimension contributes to the mass of the effective ten-dimensional particle as in $m = p$. What is the radius $\bar{R}$ of this eleventh dimension? How does it behave as a function

of g? This result is one piece of evidence for the fact that eleven-dimensional M-theory compactified on a circle is type IIA superstring theory.

Problem 24.5 Graviton and dilaton states in covariant quantization.

Examine the closed string states $\xi_{\mu\nu}\,\alpha^{\mu}_{-1}\bar{\alpha}^{\nu}_{-1}|p\rangle$, with $\xi_{\mu\nu} = \xi_{\nu\mu}$.

(a) Show that the Virasoro constraints give the conditions $p^2 = 0$ and $p_\mu \xi^{\mu\nu} = 0$.
(b) Exhibit null states that generate the physical state equivalences $\xi^{\mu\nu} \sim \xi^{\mu\nu} + p^{\mu}\epsilon^{\nu} + p^{\nu}\epsilon^{\mu}$, which hold for $p^2 = 0$ and $p \cdot \epsilon = 0$.
(c) Show that there are $(D-2)(D-1)/2$ independent physical degrees of freedom in $\xi_{\mu\nu}\,\alpha^{\mu}_{-1}\bar{\alpha}^{\nu}_{-1}|p\rangle$ for each value of p_μ which satisfies $p^2 = 0$. These are the degrees of freedom of a graviton and a scalar dilaton.

25 String interactions and Riemann surfaces

The world-sheets of interacting open strings are recognized to be Riemann surfaces, and interaction processes are seen to construct the moduli spaces of these surfaces. Conformal mapping is used to provide canonical presentations for interacting light-cone world-sheets. The celebrated Veneziano amplitude for the interaction of open string tachyons is motivated and discussed.

25.1 Introduction

Interactions and the forces that mediate them make the world interesting. If the electron and the proton did not interact, there would be no hydrogen atom. The fine structure constant $\alpha = e^2/(4\pi\hbar c)$ quantifies the strength of electromagnetic interactions and determines the interaction potential between the electron and the proton (see Section 13.4). The hot filament of a light bulb emits photons, some of which are absorbed by your eye. Emission and absorption processes are also interactions. A neutron can turn into a proton, an electron, and an antineutrino. This process, called β-decay, is the result of a weak interaction.

In string theory the strength of interactions is parameterized by the string coupling g. The value of this dimensionless number is determined by the expectation value of the dilaton field, as we discussed in Section 13.4. The string coupling g, together with the slope parameter α', determines the value of Newton's constant. The constants g and α' also determine the tension of D-branes.

Interactions arise very elegantly in string theory because they are described by processes in which strings join and split. In these processes, the world-sheets of free strings combine to form a single world-sheet, which represents the interaction. Recall that world-sheets define the two-dimensional parameter space (τ, σ) for string propagation. For a free open string, this parameter space is an infinite strip; for a free closed string, it is an infinite cylinder.

Riemann surfaces are some of the most interesting two-dimensional surfaces. They are, roughly speaking, surfaces where the two coordinates make up a complex variable. Riemann surfaces are preserved by conformal maps, so they have no *a priori* concept of distance! Inequivalent Riemann surfaces can be distinguished by parameters called moduli. Finding a way to construct all Riemann surfaces with their associated moduli is a difficult mathematical problem. Our study of string interactions will give very strong evidence for a remarkable statement: the world-sheets of interacting strings give precisely this construction.

String world-sheets can be used to calculate quantum mechanical amplitudes for scattering. Our Riemann surface analysis will motivate the amplitude for the scattering of four tachyons, the famous Veneziano amplitude. String theory began with Veneziano's discovery of this amplitude. By reaching this point in our course, we have essentially traversed the historical path backwards!

One word about terminology before getting started. We defined the world-sheet as the spacetime surface $X^\mu(\tau, \sigma)$ which represents the history of a string. As we mentioned in the paragraph above equation (6.29), physicists sometimes use the term world-sheet to refer, in addition, to the parameter space (τ, σ). In this chapter we will use this alternative meaning of the term.

25.2 Interactions and observables

Before we begin our study of string interactions and Riemann surfaces, let us set the stage for the concept of interactions. The goal is to turn a picture of a set of interacting strings into a number which gives the probability for that event to occur. Doing so involves three steps.

(1) Drawing the string diagram and calculating the conformal map which gives a canonical representation of that diagram.
(2) Using conformal field theory to calculate the scattering amplitude from the canonical representation.
(3) Using formulae to turn the scattering amplitude into a cross section. The cross section is observable.

We will study step number (1) in some detail. This entails understanding the fascinating relation between string diagrams and Riemann surfaces. Step number (2), in all its generality, is the subject of more advanced string theory courses. Step (3) properly belongs to a quantum field theory course. Step (1) is in many ways the most nontrivial of the three.

Let us examine these three steps in the case of point particles in four-dimensional spacetime. We consider a specific theory with two types of scalar particles: ϕ particles with mass M and χ particles with mass m. Suppose that $M > 2m$, so that a ϕ particle can decay into two χ particles. Indeed, in the rest frame of the ϕ particle the total energy is Mc^2, and to create two χ particles we need at least $2mc^2$. Thus $M > 2m$ is required by energy conservation, and the difference $(M - 2m)c^2$ goes into the kinetic energy of the resulting χ particles.

The diagram of step (1) is simply a picture (shown in Figure 25.1). In this picture, called a Feynman diagram, three lines come together at a point. One of the lines, the dotted one, represents the ϕ particle, and the other two represent the χ particles. For step (2) we must calculate the decay amplitude. In order to do this, we introduce a constant λ which quantifies the strength of the interaction. The value of λ is a parameter in the theory. If we

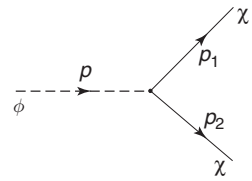

Fig. 25.1 The diagram showing the interaction of a ϕ particle with two χ particles.

let p denote the momentum of the incoming ϕ particle and p_1, p_2, denote the momenta of the outgoing χ particles, then the amplitude for an initial state consisting of a ϕ particle to turn into a final state consisting of the two outgoing χ particles is

$$_{\text{out}}\langle p_1, p_2 \mid p\rangle_{\text{in}} = -i\,(2\pi)^4\delta(p - p_1 - p_2)\,T(p, p_1, p_2), \tag{25.1}$$

where the decay amplitude T is given by

$$T(p, p_1, p_2) = \lambda. \tag{25.2}$$

The four-dimensional delta function $\delta(p - p_1 - p_2)$ ensures that both energy and momentum are conserved. Step (3) is the construction of an observable. In this case we calculate the lifetime τ_ϕ of the ϕ particle. Converting the amplitude into a lifetime is a standard calculation. The answer is

$$\frac{1}{\tau_\phi} = \frac{\lambda^2}{32\pi}\frac{1}{M}\sqrt{1 - \frac{4m^2}{M^2}}. \tag{25.3}$$

This is reasonable: as the interaction parameter λ goes to zero the ϕ particle lifetime becomes infinitely long. The factors multiplying λ^2 are kinematical; they take into account the state space available to the decay particles.

Things get more interesting when we consider the possibility of two χ particles scattering off each other. In that case both the initial and final states consist of two χ particles. We call p_1 and p_2 the momenta of the incoming particles and p_1' and p_2' the momenta of the outgoing particles. We now construct the simplest Feynman diagrams that can represent this scattering process. Since two χ can turn into a ϕ, and a ϕ can then turn into two χ, the scattering process will use the basic $\phi\chi\chi$ interaction twice. There are three possible Feynman diagrams for this scattering process, as shown in Figure 25.2. Feynman diagrams are a useful visual tool. In the path-integral formulation of quantum mechanics, the amplitude for a process is obtained by adding the amplitudes of all the "paths" consistent with the initial and final states. A Feynman diagram for a process is a representation of a specific class of allowed paths.

In diagram (a), the incoming χ with momenta p_1 and p_2 join to form a ϕ particle with momentum $p_1 + p_2$. This particle then decays into two χ with momenta p_1' and p_2'. In this process, all external particles – all the χ – are on the mass-shell. That is, they all satisfy $p^2 = -m^2$ and thus represent physical particle states. How about the intermediate ϕ particle? The momentum of this particle is $p_1 + p_2$, and, generally, $(p_1 + p_2)^2 \neq -M^2$.

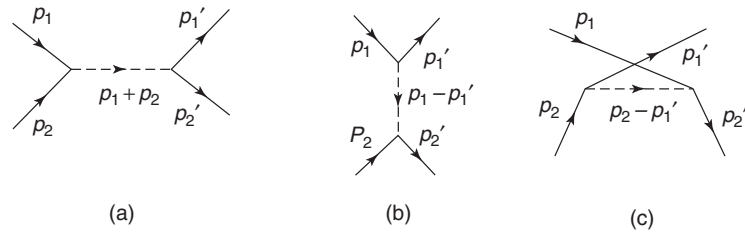

Fig. 25.2 **The three possible Feynman diagrams for the scattering of two χ particles. In each diagram the intermediate particle is a ϕ particle.**

The intermediate ϕ particle is typically *not* on the mass-shell; the intermediate states are not physical particle states. If the initial state is carefully prepared, ϕ may be on the mass-shell: the center-of-mass energy available would have to coincide precisely with Mc^2. Diagram (a) contributes to the scattering amplitude whether or not the ϕ particle is on the mass-shell.

Since all we measure are initial and final states, the interactions can occur in a more subtle way. In particular, there are diagrams where the intermediate ϕ particle could *never* be on the mass-shell! Consider diagram (b), where the incoming χ with momentum p_1 turns into the outgoing χ with momentum p_1' plus an intermediate ϕ. This intermediate ϕ could not possibly be physical – energy conservation rules out the production of physical χ and ϕ particles from a single physical χ particle (why?). The intermediate particle joins the incoming χ particle with momentum p_2 to give the outgoing χ with momentum p_2'. Diagram (c) shows the last possibility: the χ with momentum p_2 creates an intermediate ϕ and the outgoing χ with momentum p_1'. The intermediate particle joins the χ particle with momentum p_1 to give the outgoing χ with momentum p_2'. In these three diagrams, the intermediate momentum carried by the ϕ particle is given by

$$\begin{aligned} \text{first diagram}: &\quad p_1 + p_2, \\ \text{second diagram}: &\quad p_1 - p_1', \\ \text{third diagram}: &\quad p_2 - p_1'. \end{aligned} \tag{25.4}$$

This completes step (1).

Given the three diagrams, we can now proceed to step (2), the calculation of the scattering amplitude. This is the analog of equations (25.1) and (25.2). The Feynman rules of field theory give

$$_{\text{out}}\langle\, p_1', p_2' \mid p_1, p_2 \,\rangle_{\text{in}} = -i\,(2\pi)^4 \delta(p_1 + p_2 - p_1' - p_2')\, T(p_1, p_2; p_1', p_2'), \tag{25.5}$$

where

$$T(p_1, p_2; p_1', p_2') = -\lambda^2 \Big(\frac{1}{(p_1 + p_2)^2 + M^2} + \frac{1}{(p_1 - p_1')^2 + M^2} + \frac{1}{(p_2 - p_1')^2 + M^2} \Big). \tag{25.6}$$

It is useful to define invariant quantities s, t, and u that represent the intermediate momenta squared. We let

$$s = -(p_1 + p_2)^2, \quad t = -(p_1 - p_1')^2, \quad u = -(p_2 - p_1')^2. \tag{25.7}$$

In the center-of-mass frame the total spatial momentum is zero, so $p_1 + p_2 = (E_1 + E_2, \vec{0})$, where E_1 and E_2 are the center-of-mass energies of the incoming particles. Since s is a Lorentz invariant, it can be evaluated in any frame. In the center-of-mass frame we find $s = (E_1 + E_2)^2$. The invariant s thus gives the square of the total energy available in the center-of-mass frame. Using the invariants s, t, and u the scattering amplitude is rewritten as

$$T(s,t,u) = \lambda^2 \left(\frac{1}{s - M^2} + \frac{1}{t - M^2} + \frac{1}{u - M^2} \right). \tag{25.8}$$

Each denominator is of the form $(p^2 + M^2)$, where p is the momentum carried by the intermediate particle. When the intermediate particle is on the mass-shell, the denominators vanish, and we get a pole in the amplitude. In this way, the poles of scattering amplitudes tell us about particles in the theory: the corresponding values of $-p^2$ give the particle masses. Thus, for example, if we only knew about the χ particles of mass m, but were given the scattering amplitude above, we could deduce that there is an intermediate particle with mass M. This completes step (2).

The scattering amplitude, while very interesting, is not an observable. It is used to calculate the scattering cross section σ, a quantity with the dimension of area. In the problem we considered above, the cross section is the effective area that a target χ particle presents to an incoming beam of χ particles. If the incoming beam has a cross sectional area A, the probability $\mathcal{P}$ that any of the beam particles will interact with the target is σ/A. Total cross sections can be refined into differential cross sections $d\sigma$, giving, for example, the effective area for scattering particles into a small solid angle $d\Omega$ in some specified direction. The differential cross section is the typical quantity that theorists must calculate and experimentalists must measure. The Standard Model has largely been tested by comparing predicted cross sections with measured cross sections.

Differential cross sections are calculated using field theory methods. They are integrals of $|T|^2$ (times suitable kinematic factors) over the available phase space. For the example we have considered, the result is

$$\frac{d\sigma}{d\Omega} = \frac{1}{4s} \left| \frac{T}{4\pi} \right|^2. \tag{25.9}$$

This completes step (3).

25.3 String interactions and global world-sheets

Strings can interact in a rather limited number of ways. Two possible interactions of open strings go as follows: an open string can split into two open strings, and two open strings

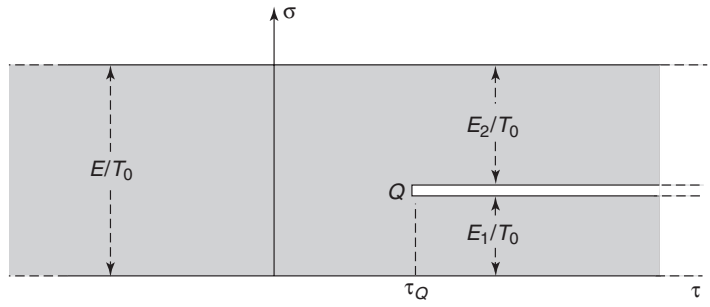

Fig. 25.3 **The basic open string interaction, in which a string of energy E splits into strings of energies E_1 and E_2. The interaction occurs at $\tau = \tau_Q$.**

can join to form a single open string. String diagrams are representations of the string interactions; they are analogous to the Feynman diagrams of particle physics. The focus of the string diagram is the (τ, σ) world-sheet. The string diagram does not represent the spacetime surface traced by the string. In a string diagram the world-sheets of free strings are put together to form the world-sheet of an interaction process.

Interactions must be analyzed using a specific gauge, since gauge choices affect the world-sheet description. Let us recall how world-sheets look in the static gauge. In this gauge

$$X^0(\tau, \sigma) = c\tau \,, \tag{25.10}$$

and strings are parameterized using the energy E that they carry:

$$\sigma \in \left[0, E/T_0\right] = [\,0, 2\pi\alpha' E\,], \tag{25.11}$$

where T_0 is the string tension. Since energy must be conserved and the energy of a string determines its σ length, the interactions must conserve the total σ length. In Figure 25.3 we show a string of energy E splitting into two strings with energies E_1 and E_2. Note that the incoming string with energy E is semi-infinite towards the past, while the outgoing strings are semi-infinite towards the future. In the far past $\tau \to -\infty$ there is only one string; in the far future $\tau \to +\infty$ there are two strings. The interaction point Q is also shown. It occurs at some value τ_Q of τ. At each $\tau < \tau_Q$ there is just one string, and at each $\tau > \tau_Q$ there are two strings.

Somewhat more complicated processes are also possible. Two strings, with energies E_1 and E_2, can come together to form a single open string, of energy $E_1 + E_2$. This intermediate string can then split into two open strings, of energies E_3 and E_4 with $E_3 + E_4 = E_1 + E_2$ (see Figure 25.4). This string diagram can be used to calculate a string amplitude regardless of the types of particles that are represented by the quantum states of the strings. For example, the initial quantum states of the two incoming strings could be two tachyons or two photons or a photon and a tachyon or any other two particle states in the quantum string spectrum. The same is true for the outgoing states. The string diagrams are universal; the information about the specific particles only enters in step (2) of the amplitude calculation.

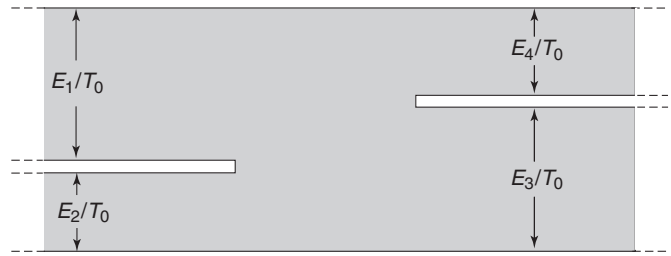

Fig. 25.4 **Incoming open strings one and two join to form an intermediate open string that then splits into strings three and four.**

We now consider string diagrams in the light-cone gauge. In this gauge we have ((9.62) with $\beta = 2$)

$$X^+ = 2\alpha' p^+ \tau, \quad p^+ = \pi \mathcal{P}^{\tau +}. \tag{25.12}$$

The second equation, in particular, implies that $\sigma \in [0, \pi]$ for every string. The above equations are not suitable to deal with interactions. If every string has $\sigma \in [0, \pi]$, then σ length cannot be conserved in interactions, and it is not clear anymore how to draw string diagrams! There is an additional difficulty. Imagine that we have two open strings (strings one and two) which interact to form a single open string (string three). We select a Lorentz frame, use light-cone coordinates, and, with the above version of the light-cone gauge choice applied to each of the three strings, we have

$$\begin{aligned} X_1^+ &= 2\alpha' p_1^+ \tau_1, \\ X_2^+ &= 2\alpha' p_2^+ \tau_2, \\ X_3^+ &= 2\alpha' p_3^+ \tau_3 . \end{aligned} \tag{25.13}$$

Here the subscripts label the strings. The interaction is a particular event in spacetime, and the three strings must agree on the value of the physical time X^+ at which it occurs. Moreover, since we are to combine the three world-sheets into a single world-sheet, with a single τ coordinate, the interaction point must have a unique τ. This value of τ is generically nonzero, so the above equations, with different values of p_i^+, will in general not give $X_1^+ = X_2^+ = X_3^+$.

How can we modify our light-cone setup in order to deal with interactions? Recall the more general version of equation (9.62), which applies for arbitrary β (see the remarks below (9.34)):

$$X^+ = \beta\alpha' p^+ \tau, \quad p^+ = \frac{2\pi}{\beta} \mathcal{P}_\mu^+ . \tag{25.14}$$

Formerly, the choice $\beta = 2$ was taken in order to get open strings with $\sigma \in [0, \pi]$. The idea now is to choose β so as to eliminate the p^+ dependence from the relation between X^+ and τ. If we choose

$$\beta = \frac{1}{\alpha' p^+}, \tag{25.15}$$

then we will have the gauge

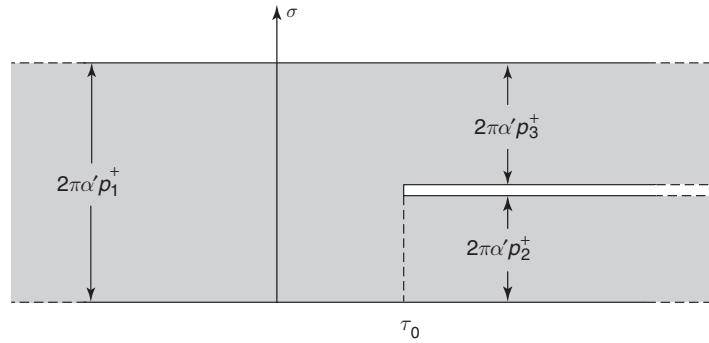

Fig. 25.5 **The global world-sheet for a splitting interaction in the light-cone gauge. The conservation of p^+ implies that the total width of the strip is constant in τ.**

$$X^+ = \tau, \tag{25.16}$$

which makes it possible to use a single τ for all interacting strings. For the case of the three strings just discussed, if we call τ_0 the interaction time, then we have

$$\begin{aligned} X_1^+ &= \tau, \quad \tau \le \tau_0, \\ X_2^+ &= \tau, \quad \tau \ge \tau_0, \\ X_3^+ &= \tau, \quad \tau \ge \tau_0. \end{aligned} \tag{25.17}$$

String one ceases to exist at τ_0, at which time strings two and three are born. Note that $X_2^+ = X_3^+$ when $\tau \ge \tau_0$. At $\tau = \tau_0$ the three X^+ agree. Along with (25.16), each string is parameterized by

$$\sigma \in \left[\, 0, \frac{2\pi}{\beta} \,\right] = [\, 0, 2\pi\alpha' p^+], \tag{25.18}$$

where p^+ is the momentum carried by the string. Had we used this convention to quantize the string in our earlier work, unwieldy factors would have appeared in many of the equations.

Now we can represent the splitting of string one into strings two and three using a single diagram in the (τ, σ) plane, as shown in Figure 25.5. The interaction looks like a constant width infinite strip with a cut. The figure is described by two parameters: the width of string one and the width of either string two or string three. Where we start the cut on an infinite strip is just a matter of convention, so the interaction time τ_0 is not really a parameter.

25.4 World-sheets as Riemann surfaces

We will now show that it is reasonable to view string world-sheets as Riemann surfaces. Riemann surfaces are two-dimensional surfaces. The simplest Riemann surface is the complex plane $\mathbb{C}$. This is the conventional (x, y) plane, but with the understanding that the

complex coordinate $z \equiv x + iy$ is used to describe points on the plane. Subregions of the complex plane are also Riemann surfaces. For example, the upper half-plane $\mathbb{H}$, defined to include all points z with positive imaginary part, $\Im(z) > 0$, is an important Riemann surface. A closely related surface is the bordered upper half-plane $\bar{\mathbb{H}}$, which includes all points z with $\Im(z) \geq 0$ and the "point at infinity" (which is discussed in Section 25.6). An annulus is also a Riemann surface. More complicated Riemann surfaces include the Riemann sphere $\hat{\mathbb{C}}$ and the torus. We will have the chance to study all of these in some detail, either in this chapter or in the following one.

In this and in several of the following sections the level of mathematical rigor will be higher than before. Some familiarity with complex variables will be helpful. The definition of a Riemann surface (to be given below) is included for completeness and will not be strictly necessary for the material that follows.

The definition of a Riemann surface requires the concept of an analytic function. A function $f(z)$ of the complex variable z is analytic at z_0 if the derivative $f'(z)$ exists in some neighborhood of z_0. If we write $f(z) = u(x, y) + iv(x, y)$, then the functions u and v satisfy the Cauchy–Riemann equations

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}, \tag{25.19}$$

over the domain of analyticity. A Riemann surface is a two-dimensional real manifold equipped with complex charts: homeomorphisms z_α which take open sets U_α into open subsets of the complex plane $\mathbb{C}$ (a homeomorphism is a one-to-one continuous map with a continuous inverse). Here α is a label for the open sets U_α that cover the manifold. Over the intersections $U_\alpha \cap U_\beta$, the transition functions $z_\alpha \circ z_\beta^{-1}$ must be analytic.

Two Riemann surfaces are considered equivalent if there is a continuous one-to-one and onto mapping relating them that is analytic. This means that the map, expressed in terms of the charts, is an analytic function from a subset of $\mathbb{C}$ to some subset of $\mathbb{C}$. An analytic map is *conformal*: angles between arcs meeting at a point are preserved under the map, except possibly at points where the derivative of the map vanishes (we will see examples of this soon). Even though angles are preserved locally, the overall shape of the surface can change quite dramatically under a conformal map. In this and in the next chapter we will take conformal to mean analytic and one-to-one. Accordingly, the existence of a conformal map taking one Riemann surface onto another Riemann surface implies the equivalence of the two Riemann surfaces.

Since conformally related surfaces are the same Riemann surface, the concept of distance cannot be defined intrinsically. Consider the complex z-plane and two points P_1 and P_2, with coordinates z_1 and z_2, respectively. The natural definition of distance between two points in $\mathbb{C}$ is given by the absolute value $|z_1 - z_2|$ of the difference of coordinates. Now map the z-plane into the w-plane via $w = 2z$. The two planes are the same Riemann surface. After the conformal map, the images of P_1 and P_2 (still called P_1 and P_2 in the w-plane), have coordinates $w_1 = 2z_1$, and $w_2 = 2z_2$. The distance between P_1 and P_2 in the w-plane is thus $|w_1 - w_2| = 2|z_1 - z_2|$. The distance is doubled. Since both planes are the

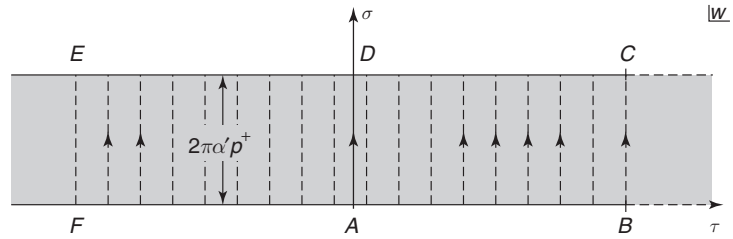

Fig. 25.6 **An open string freely propagating from $\tau = -\infty$ to $\tau = +\infty$. The string diagram is a strip of constant width $2\pi\alpha' p^+$. The vertical dashed segments are the open strings.**

same Riemann surface, we are at a loss when asked what is the distance between the points P_1 and P_2. Conformal maps locally scale distances. In fact, locally, lengths measured in all directions are scaled by the same factor. From $w = f(z)$ we find $dw = f'(z)dz$. It follows that at any point z_0 we have $|dw| = |f'(z_0)||dz|$. Since $|dw|$ is the length of the image of the vector dz, we see that, locally, all lengths are scaled by the same factor $|f'(z_0)|$.

Perhaps this reminds you of a metric whose scale turned out to be irrelevant. Indeed, a world-sheet metric $h_{\alpha\beta}$ enters in the Polyakov action (24.70), and Weyl transformations $h_{\alpha\beta} \to \Omega^2 h_{\alpha\beta}$ leave the string action invariant (see (24.91)). Weyl transformations scale all metric components by the same factor at each point on the world-sheet. As a result, locally, all lengths are scaled in the same way, just as a conformal map does for Riemann surfaces. The invariance of the string action under local scale transformations of the world-sheet metric indicates that the scale of lengths on the world-sheet is not physical. Since the scale of lengths is not well defined on Riemann surfaces, this suggests that viewing world-sheets as Riemann surfaces is natural.

While the scale of distances behaves similarly, there is one difference. The world-sheet metric can be set to be proportional to the two-dimensional Minkowski metric $\eta_{\alpha\beta}$, which defines distances as $-ds^2 = -d\tau^2 + d\sigma^2$. On the other hand, on the complex plane the natural metric is Euclidean: $ds^2 = |dz|^2 = dx^2 + dy^2$. Because of this sign difference, we cannot really prove that world-sheets can be treated as Riemann surfaces. But nothing stops us from trying to treat them as such, especially because it is known that much about a Minkowski theory can be learned from its Euclidean version. It turns out that thinking of world-sheets as Riemann surfaces leads to a consistent picture of string interactions.

In order to view world-sheets as Riemann surfaces, our first task is to find suitable complex coordinates. We assemble τ and σ into a complex coordinate w as follows:

$$w = \tau + i\sigma. \tag{25.20}$$

The simplest world-sheet is that of a freely propagating open string. As shown in Figure 25.6, the world-sheet for a string with light-cone momentum p^+ is the strip $0 \le \Im(w) \le 2\pi\alpha' p^+$ ($\Im(w)$ denotes the imaginary part of w and $\Re(w)$ denotes the real part of w). The strings are the vertical segments of constant τ. Since conformal maps do

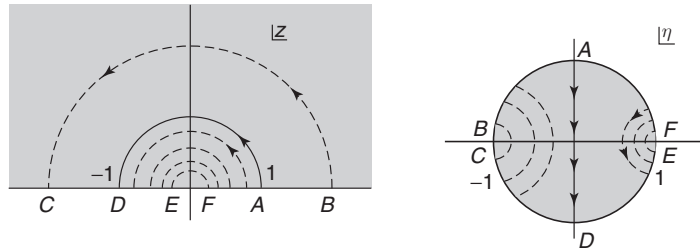

Fig. 25.7 **Left: the infinite strip is mapped into $\bar{\mathbb{H}}$ via the exponential map. The infinite past is at $z = 0$, and the infinite future is at $z = \infty$. The open strings are the dashed lines. Right: $\bar{\mathbb{H}}$ is mapped onto the unit η disk via a linear fractional transformation. The infinite past is at $\eta = 1$, and the infinite future is at $\eta = -1$.**

not change the physics, it is a good idea to familiarize ourselves with pictures that are conformally equivalent to the strip.

Let us see what happens to the strip when we map it conformally using the exponential function

$$z = \exp\left(\frac{w}{2\alpha' p^+}\right). \tag{25.21}$$

(Recall that $\exp(u + iv) = \exp(u)\,(\cos v + i \sin v)$.) The result of this mapping is shown in the left part of Figure 25.7. Note that the full string in the infinite past limit is mapped to a point! Indeed, for $w = -\tau_0 + i\sigma$, with $\tau_0 \to \infty$ we have

$$z = \exp\Big(-\frac{\tau_0}{2\alpha' p^+}\Big)\exp\Big(i\,\frac{\sigma}{2\alpha' p^+}\Big) \longrightarrow 0, \tag{25.22}$$

for all values of σ. As shown in the figure, the strings that approach the infinite past approach the point $z = 0$. Note also another curious fact. The two boundaries of the strip are mapped to the real line. The boundary $\sigma = 0$ is mapped to the positive half of the real line, while the boundary $\sigma = 2\pi\alpha' p^+$ is mapped to the negative half of the real line. The angle between the boundaries at the infinite past has been changed. Think of the boundaries as lines emerging from the past. While such boundaries make zero angle in the strip picture, they make an angle of 180° in the z-picture. The strip has been mapped into $\bar{\mathbb{H}}$, and the strings appear as semicircles centered at $z = 0$. The string at $\tau = 0$ is the unit semicircle. The strings in the far future are very large semicircles on the z-plane. They grow without bound as the strings approach the infinite future.

There is another useful presentation for the Riemann surface of a free open string. We can map $\bar{\mathbb{H}}$ onto the unit disk in the η plane via the transformation

$$\eta = \frac{1 + iz}{1 - iz}. \tag{25.23}$$

This mapping, illustrated in the right side of Figure 25.7, takes the real line $\Im(z) = 0$ into the unit circle. Indeed, for $z = x \in \mathbb{R}$ we have

$$|\eta|^2 = \eta\eta^* = \frac{1+ix}{1-ix} \cdot \frac{1-ix}{1+ix} = 1. \tag{25.24}$$

This time the infinite past $z = 0$ is mapped to $\eta = 1$, and the infinite future $z = \infty$ is mapped to $\eta = -1$. The two boundary components are the unit semicircles BAF and CDE.

We must draw one important lesson from the above maps. When the strip is mapped onto a bounded region, the complete strings in the infinite past and in the infinite future are mapped into two *points* on the boundary of the bounded region. More precisely, these points are limits that are not attained for any finite time string, so topologically these points are to be removed from the boundary of the bounded region: they are *punctures*. We say that strings are inserted at the punctures. Our conformal maps have shown that the infinite strip is conformally equivalent to a disk, twice-punctured on the boundary. The punctures break the boundary of the unit disk into two disjoint components. In the mapping of the strip to $\bar{\mathbb{H}}$, the point $z = 0$ is not attained for any string in the finite past. The far future on the strip is mapped to large z, and it will become clear in Section 25.6 that the "point at infinity" is not attained either. Therefore, the infinite strip maps onto $\bar{\mathbb{H}}$ with two points removed.

A free closed string can be represented by the string diagram in Figure 25.6, with the understanding that the top and bottom edges of the infinite strip are identified. More precisely, we identify points on the two edges that have the same value of $\Re(\tau)$. This identification results in an infinite cylinder of circumference $2\pi\alpha' p^+$. The vertical dashed segments become closed strings.

Quick calculation 25.1 Show that

$$z = \exp\Big(\frac{w}{\alpha' p^+}\Big) \tag{25.25}$$

maps the free closed string world-sheet to the full z-plane with the origin removed. Verify that the vertical dashed segments that represent closed strings in the w-plane become closed circles on the z-plane.

Note that under the map (25.25) the closed string in the infinite past is sent to $z = 0$, while the closed strings that approach the infinite future are mapped to larger and larger circles centered at the origin. The closed string world-sheet is the complex plane punctured at the origin.

25.5 Schwarz–Christoffel map and three-string interaction

Now that we have examined some conformal maps of the world-sheet of a freely propagating open string, we turn to the case of interacting strings. The simplest interaction is one where an open string breaks into two open strings, as shown in Figure 25.5. Following

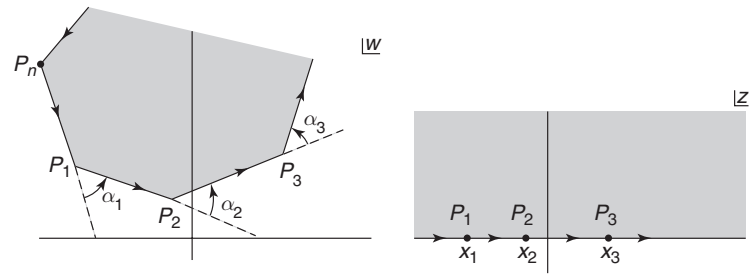

Fig. 25.8 **The Schwarz–Christoffel map $w(z)$ takes $\bar{\mathbb{H}}$ onto the polygon. Shown are the vertices P_i, and the turning angles α_i. In the (bordered) upper half-plane, shown to the right, the x_i are mapped to the vertices of the polygon.**

the intuition developed above, this surface should be conformally equivalent to a disk with three punctures on the boundary. In fact, the mapping theorem of Riemann guarantees that this interacting world-sheet can be conformally mapped into $\bar{\mathbb{H}}$, with the open string boundaries mapping into the real line.

Why do we want to map the world-sheet to $\bar{\mathbb{H}}$? $\bar{\mathbb{H}}$ gives us a canonical presentation that will allow us to compare different Riemann surfaces. This comparison will play a major role in the study of interactions that involve more than three strings. In mapping the string diagram to $\bar{\mathbb{H}}$ we fulfill step (1) of the list indicated at the beginning of Section 25.2.

It is not so easy to map the world-sheet in Figure 25.5 onto $\bar{\mathbb{H}}$. This world-sheet, however, can be viewed as the limit of a polygon, and there is a well established method for constructing the map that takes a polygon onto $\bar{\mathbb{H}}$. The map is conformal, except at the corners of the polygon. In general, one cannot write the conformal map explicitly, but one can always write a differential equation for the mapping function. The maps relating polygons and $\bar{\mathbb{H}}$ are called Schwarz–Christoffel maps.

Suppose we have a polygon in the complex w-plane. Let the polygon have n sides and therefore n vertices, denoted by $P_1, P_2, \ldots, P_n$. The Schwarz–Christoffel map $w(z)$ takes $z \in \bar{\mathbb{H}}$ into the polygon. The situation is illustrated in Figure 25.8. We choose the boundary of the polygon to be oriented so that the interior of the polygon lies to the left of the oriented boundary. The mapping takes the real line in the z-plane, oriented in the direction of increasing values, into the oriented polygon boundary.

The polygon has turning angles α_i at the vertices P_i. A turning angle is positive if the oriented side emerging from the vertex is obtained by counterclockwise rotation from the extension of the side that enters the vertex. By definition, we restrict turning angles to the interval $[-\pi, \pi]$

$$-\pi \leq \alpha_i \leq \pi. \tag{25.26}$$

A turning angle of value near to but smaller than π represents a corner where the polygon covers only a small sector. On the other hand, a turning angle of value near to but larger than $-\pi$ represents a corner where the polygon fails to cover a small sector. Turning

angles of value π or $-\pi$ give degenerate polygons. In fact light-cone string diagrams are degenerate polygons. The angles π and $-\pi$ give completely different types of corners.

We claim that the differential equation for the mapping function $w(z)$ is

$$\frac{dw}{dz} = A(z - x_1)^{-\frac{\alpha_1}{\pi}}(z - x_2)^{-\frac{\alpha_2}{\pi}} \cdots (z - x_{n-1})^{-\frac{\alpha_{n-1}}{\pi}}. \tag{25.27}$$

Here the x_i are a collection of ordered real numbers:

$$x_1 < x_2 < \cdots < x_{n-1}, \tag{25.28}$$

and the map $z(w)$ takes the vertices to those points:

$$z(P_i) = x_i, \quad i = 1, 2, \ldots, n-1. \tag{25.29}$$

Note that the turning angle α_n does not appear in the differential equation (25.27). In fact, the map constructed from (25.27) will take P_n to $z = \infty$:

$$z(P_n) = \infty. \tag{25.30}$$

Since a polygon is closed, the turning angles must sum to 2π:

$$\alpha_1 + \alpha_2 + \cdots + \alpha_n = 2\pi. \tag{25.31}$$

Thus α_n is determined by the other turning angles. This is why the map in (25.27) can properly handle the "missing" vertex (see Problem 25.6 for details).

Now we will explain why (25.27) is the correct differential equation for the Schwarz–Christoffel map. We do so by taking the argument of both sides of the equation. The argument of a complex number $z = r\exp(i\theta)$, with r real, is defined as $\arg(z) = \theta$. Since angles are multivalued, the arg function has an additive ambiguity, we can add or subtract any multiple of 2π. The function arg satisfies $\arg(z_1 z_2) = \arg(z_1) + \arg(z_2)$ and $\arg(z_1/z_2) = \arg(z_1) - \arg(z_2)$. Taking the argument of both sides of (25.27) and using these properties, we find

$$\arg(dw) - \arg(dz) = \arg A + \arg(z - x_1)^{-\frac{\alpha_1}{\pi}} + \cdots + \arg(z - x_{n-1})^{-\frac{\alpha_{n-1}}{\pi}}. \tag{25.32}$$

We fix our conventions by replacing the arg functions on the right-hand side with the principal value Arg, which satisfies

$$-\pi < \mathrm{Arg}(z) \leq \pi. \tag{25.33}$$

Arg of a positive real number is zero, while Arg of a negative real number is π. If we define z^β as the principal branch $z^\beta \equiv \exp[\,\beta\,(\ln|z| + i\,\mathrm{Arg}(z))\,]$, we then have

$$\mathrm{Arg}(z^\beta) = \beta\,\mathrm{Arg}(z), \quad \text{for} \quad |\beta| \leq 1. \tag{25.34}$$

The condition $|\beta| \leq 1$ guarantees that the right-hand side satisfies the inequalities appropriate for Arg of a complex number. Since $|\alpha_i/\pi| \leq 1$, we can use (25.34) to simplify (25.32):

$$\arg(dw) - \arg(dz) = \mathrm{Arg}\,A - \frac{\alpha_1}{\pi}\mathrm{Arg}(z - x_1) - \cdots - \frac{\alpha_{n-1}}{\pi}\mathrm{Arg}(z - x_{n-1}). \tag{25.35}$$

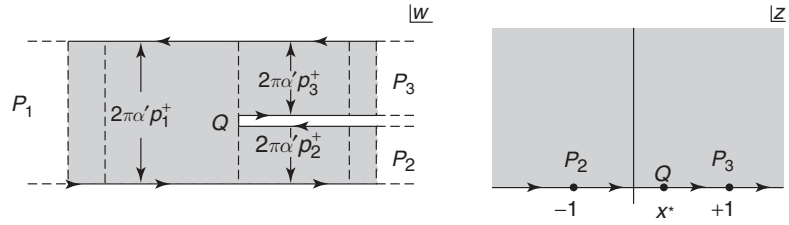

Fig. 25.9 **The mapping of a three-string interaction into the upper half-plane. The string diagram is a degenerate polygon with vertices P_1, P_2, Q, and P_3. The interaction point Q is mapped to the point $z = x^*$. The vertex P_1 is mapped to infinity.**

When $z = x$ is on the real line and $dz = dx > 0$, equation (25.35) gives

$$\arg(dw) = \mathrm{Arg}A - \frac{\alpha_1}{\pi}\mathrm{Arg}(x - x_1) - \cdots - \frac{\alpha_{n-1}}{\pi}\mathrm{Arg}(x - x_{n-1})\,. \tag{25.36}$$

Since the real line in the z-plane maps to the boundary of the polygon, $\arg(dw)$ is the angle that an edge makes with respect to the horizontal axis. As long as x varies without crossing a turning point x_i, nothing changes on the right-hand side: the arguments of the positive quantities remain equal to zero, and the arguments of the negative quantities remain equal to π. This generates a straight edge in the w-plane. On the other hand, as x goes from the left side to the right side of a turning point x_i, $\mathrm{Arg}(x - x_i)$ changes from π to 0, a total change of $(-\pi)$. The right-hand side of the above equation then changes by $(-\alpha_i/\pi)(-\pi) = \alpha_i$. This means that dw along the new edge will have an argument larger by α_i, which is exactly what is supposed to happen. This confirms our claim that the differential equation does its job. Given a specific polygon, the turning points x_i as well as the constant A have to be calculated. These quantities depend on the length of the sides of the polygon. We will first apply the Schwarz–Christoffel transformation to the case of an open string splitting in two, as shown in Figure 25.9. This world-sheet can be viewed as a degenerate polygon with four vertices. If we orient the boundary of the polygon starting with the lower horizontal line, the first vertex we find is P_2, the second string in the infinite future. Here we have a turning angle of $+\pi$. The next edge travels back to the interaction point Q. This is the second vertex, and here the turning angle is $-\pi$. The next edge goes to the infinite future P_3 of the third string. Here we find our third vertex, where we have a turning angle of $+\pi$. Finally, the fourth vertex is P_1, the infinite past of string one. Here the turning angle is $+\pi$. In summary, we have

$$\begin{aligned}
&\text{turning angle at } P_2 = \alpha_2 \;= +\pi\,,\\
&\text{turning angle at } Q \;= \alpha_Q = -\pi,\\
&\text{turning angle at } P_3 = \alpha_3 \;= +\pi,\\
&\text{turning angle at } P_1 = \alpha_1 \;= +\pi.
\end{aligned} \tag{25.37}$$

Note that the turning angles add up to 2π, as expected from (25.31). The turning angles are perhaps easier to appreciate in Figure 25.10, where the world-sheet is shown schematically

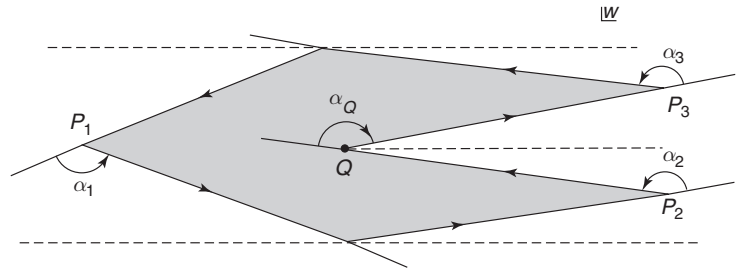

Fig. 25.10 **A nondegenerate version of the degenerate polygon shown on the left side of Figure 25.9. In the degenerate limit, the small turning angles above and below Q disappear.**

as an almost degenerate polygon by reducing the magnitude of the turning angles from the value of π and adding two small turning angles right above and right below Q.

We will map to $\bar{\mathbb{H}}$ by requiring that P_2 go to $z = -1$, P_3 go to $z = +1$, and P_1 go to $z = \infty$. Having one vertex at infinity is implicit in the differential equation we wrote, and it simplifies matters. You may ask: how do we know that we can map the other two points, P_2 and P_3, to the chosen values of z? The answer is based on a result obtained in the next section: three points on the real line can be mapped to any specific coordinates while preserving $\bar{\mathbb{H}}$ and the ordering of the points. These three points are P_2, P_3, and P_1, and we have chosen to map them to -1, $+1$, and ∞. We cannot, however, specify the coordinate x^* of the last vertex Q. As shown in the right part of Figure 25.9, x^* must lie between -1 and $+1$, since the vertex Q lies in between the vertices P_2 and P_3 of the oriented polygon.

We now proceed to write the differential equation for this map, following (25.27). With ordered turning points $z = -1, x^*, +1$ and turning angles $\alpha_2, \alpha_Q, \alpha_3$ we write

$$\frac{dw}{dz} = A(z-(-1))^{-\frac{\alpha_2}{\pi}}(z-x^*)^{-\frac{\alpha_Q}{\pi}}(z-1)^{-\frac{\alpha_3}{\pi}}. \tag{25.38}$$

Using the values for the angles from (25.37), we obtain

$$\frac{dw}{dz} = A\frac{(z-x^*)}{(z+1)(z-1)}. \tag{25.39}$$

For z real and greater than one, dw/dz must be real and negative, since the image of $z > 1$ is the horizontal line between P_3 and P_1. Therefore, we conclude that A is real and negative. In order to integrate this equation, we decompose the right-hand side using partial fractions:

$$\frac{dw}{dz} = \frac{A}{2}\left(\frac{1+x^*}{z+1} + \frac{1-x^*}{z-1}\right). \tag{25.40}$$

Integrating, we obtain

$$w = \frac{A}{2}(1+x^*)\ln(z+1) + \frac{A}{2}(1-x^*)\ln(z-1). \tag{25.41}$$

Here we define the logarithms by $\ln(z) = \ln|z| + i\,\mathrm{Arg}(z)$, with the logarithm of a real number taken to be real. We still need to determine the constants A and x^* in terms of the

parameters p_2^+ and p_3^+ of the string diagram. We need $\sigma = \Im(w)$ to increase by $2\pi\alpha' p_2^+$ as z crosses $z = -1$, the image of P_2. Since $\Im(\ln z) = \text{Arg}(z)$, and A is real, we have

$$\sigma = \Im(w) = \frac{A}{2}(1 + x^*)\text{Arg}(z + 1) + \frac{A}{2}(1 - x^*)\text{Arg}(z - 1). \tag{25.42}$$

As z goes from the left to the right of -1, $\text{Arg}(z + 1)$ goes from π to 0, and thus

$$\Delta\sigma = \frac{A}{2}(1 + x^*)(0 - (\pi)) = 2\pi\alpha' p_2^+. \tag{25.43}$$

This gives

$$\frac{A}{2}(1 + x^*) = -2\alpha' p_2^+. \tag{25.44}$$

Similarly, the behavior near $z = 1$ requires

$$\frac{A}{2}(1 - x^*) = -2\alpha' p_3^+. \tag{25.45}$$

With these two equations, back in (25.41) we have

$$w = -2\alpha' p_2^+ \ln(z + 1) - 2\alpha' p_3^+ \ln(z - 1). \tag{25.46}$$

This is the equation for the map, written in terms of known parameters.

Quick calculation 25.2 Convince yourself that the real axis of the w-plane coincides with the top edge of the strip on the left side of Figure 25.9.

In the future, we will be able to write equations like (25.46) by inspection. The rules are clear. We need a logarithm for each of the strings, except for the one whose turning point is mapped to infinity in the upper half-plane. The prefactor of each of these logarithms is related to the width of the corresponding string – it equals $(-2\alpha' p^+)$ when σ increases as we cross the turning point. We will use this understanding to write the map for a four-string interaction in Section 25.7.

The map (25.46) was written without calculating x^*. We can find x^* by dividing equations (25.44) and (25.45):

$$\frac{1 + x^*}{1 - x^*} = \frac{p_2^+}{p_3^+} \longrightarrow x^* = \frac{p_2^+ - p_3^+}{p_2^+ + p_3^+}. \tag{25.47}$$

If $p_2^+ = p_3^+$, then $x^* = 0$, as we would expect, since Q is then half-way between P_2 and P_3. If $p_2^+ \gg p_3^+$, the figure leads us to believe that, conformally speaking, Q is getting close to P_3. This is indeed the case, for then $x^* \to 1$. On the other hand, if $p_3^+ \gg p_2^+$ then $x^* \to -1$, and indeed Q is getting closer to P_2.

The value of x^* can also be calculated from (25.46). For this, note that $dw/dz = 0$ at $z = x^*$, because at Q, the derivative dw/dz goes from negative to positive (τ reaches a local minimum). Using (25.46) we find that

$$\frac{dw}{dz} = -\frac{2\alpha' p_2^+}{x^* + 1} - \frac{2\alpha' p_3^+}{x^* - 1} = 0, \tag{25.48}$$

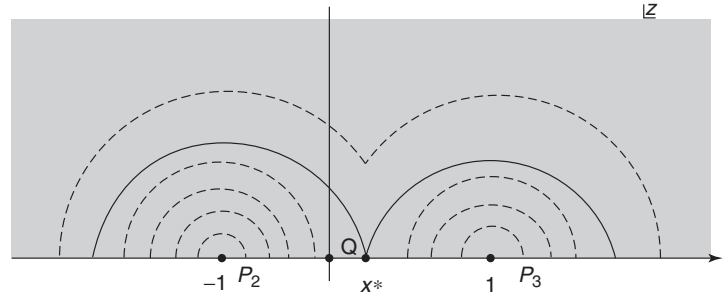

Fig. 25.11 **Three-string interaction as seen in $\bar{\mathbb{H}}$. Shown as dashed lines are the images of the (vertical) strings in Figure 25.9. The string at the interaction time is shown as a solid line.**

which leads to the same value for x^*. It is interesting to look at the strings in the z-plane, as shown in Figure 25.11. Observe the strings converging into $z=-1$ and $z=+1$. The strings arriving from P_1 at $z=\infty$ eventually reach the interaction point. The string at the interaction time is shown as a solid line. Its endpoints are to the left and to the right of P_2 and P_3, respectively, and one of its interior points touches the boundary at x^*.

Quick calculation 25.3 Consider the three-string interaction when $p_2^+ = p_3^+$, so that $x^*=0$. Check that $\Re(w(x^*))=0$. Show that the endpoints of the string at the interaction time lie at $z=\pm\sqrt{2}$.

The amplitude for the three-string interaction considered here includes the open string coupling g_o as a multiplicative factor ($g_o^2 \sim g$, where g is the closed string coupling). This is analogous to the amplitude (25.2), which also includes the coupling constant.

25.6 Moduli spaces of Riemann surfaces

Since surfaces related by a conformal map are, by definition, the same Riemann surface, it is important to understand when we can map a pair of surfaces into each other. To make this question more concrete, consider a simple example. Take the upper half-plane $\bar{\mathbb{H}}$ with one puncture at a position x on the real axis. Can this configuration be mapped to the upper half-plane $\bar{\mathbb{H}}$ with a puncture $x' \neq x$? Are these two the same Riemann surface? They are. Use the coordinate z for the first $\bar{\mathbb{H}}$ and the coordinate z' for the second $\bar{\mathbb{H}}$. Then take

$$z' = z + (x' - x). \tag{25.49}$$

This conformal map (z' is an analytic function of z) takes the puncture at $z=x$ into the puncture at $z'=x'$. This map also preserves $\bar{\mathbb{H}}$ since $\Im(z')=\Im(z)$. The upshot of this example is that $\bar{\mathbb{H}}$ with one puncture on the boundary is a Riemann surface without a parameter. Whichever point is chosen, the Riemann surface is the same. We call the

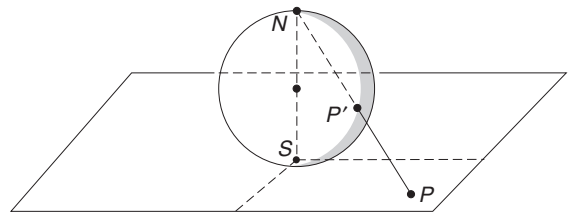

Fig. 25.12 **A map from the plane to the sphere. The sphere is resting on the plane with the south pole *S* at the origin. For each point *P* in the plane we associate the point *P′* on the sphere. The point at infinity maps to the north pole *N*.**

parameters of Riemann surfaces moduli, or modulus, if there is just one. Thus, $\bar{\mathbb{H}}$ with one puncture on the real axis has no moduli.

As we will see here, and in the exercises at the end of this chapter, the story is more interesting when we add additional punctures on the real line. $\bar{\mathbb{H}}$ with two punctures still has no moduli! The positions of the punctures can be changed at will. $\bar{\mathbb{H}}$ with three punctures still has no continuous modulus, but it does have one discrete modulus, which describes the ordering of the punctures on the boundary of $\bar{\mathbb{H}}$. Finally, $\bar{\mathbb{H}}$ with four punctures has one continuous modulus as well as a discrete one.

We begin our analysis with the complex plane $\mathbb{C}$, extended to define the Riemann sphere. We will then return to the case of $\bar{\mathbb{H}}$, which is immediately applicable to open string interactions. Our analysis of the Riemann sphere will be needed in the following chapter, where we discuss closed strings.

The complex plane $\mathbb{C}$ can be extended to construct the Riemann sphere $\hat{\mathbb{C}}$. Imagine a sphere sitting atop the complex plane, such that the south pole coincides with the origin and the north pole is directly above it (see Figure 25.12). Every point P on the complex plane can be matched to a point P' on the sphere using a stereographic projection: P' is the intersection of the sphere with the line that connects P to the north pole. Clearly, this is a one-to-one map. As we move out on the complex plane, the points on the sphere approach the north pole. The image of $\mathbb{C}$ under the map is the sphere without its north pole – the sphere minus one point. To build the complete sphere we must extend the complex plane by the inclusion of a "point at infinity," a point $\{\infty\}$ whose image under the map is declared to be the north pole. Via the stereographic map, the extended complex plane $\mathbb{C} \cup \{\infty\}$ is equivalent to the Riemann sphere $\hat{\mathbb{C}}$. This is why the Riemann sphere is sometimes called the extended complex plane. In practice, to work with a Riemann sphere, we use the complex plane with the understanding that $z = \infty$ is a point, to be treated just like every finite point.

The Riemann sphere $\hat{\mathbb{C}}$ is relevant to closed strings. Consider the map (25.25) from the infinite cylinder to the complex z-plane. There is some asymmetry here: in the cylinder picture the closed strings in the infinite past and infinite future appear symmetrically, but on the plane the first is seen as a puncture at $z = 0$, while the second is mapped to an infinite circle. To restore the symmetry the solution is clear: the infinite circle on the plane must be viewed as a single puncture. We therefore say that the world-sheet of a free closed string

is the extended complex plane $\mathbb{C} \cup \{\infty\}$ with the point $z = 0$ removed and the point at infinity removed. Equivalently, the world-sheet of a free closed string is a Riemann sphere $\hat{\mathbb{C}}$ with two punctures. The punctures are $z = 0$ and the point at infinity; both are on the same footing in $\hat{\mathbb{C}}$.

What is the most general analytic one-to-one map from $\hat{\mathbb{C}}$ to $\hat{\mathbb{C}}$? Mathematicians have shown that it is a *linear fractional transformation*. Consider two copies of $\hat{\mathbb{C}}$, with coordinates z and w. A linear fractional transformation takes the form

$$w = \frac{az + b}{cz + d}, \tag{25.50}$$

where $a, b, c,$ and d are complex numbers. Note that the point at infinity in the z-plane is mapped to $w = a/c$, which is typically a finite number. This is consistent with our view that in $\hat{\mathbb{C}}$ the point at infinity is a generic point. For the map to be one-to-one, we must, in addition, require

$$ad - bc \neq 0. \tag{25.51}$$

If $a = b = 0$, for example, this condition is violated, and (25.50) gives $w = 0$ for all z. To show that (25.51) guarantees a one-to-one map, we must show that $w(z_1) = w(z_2)$ implies $z_1 = z_2$. Using (25.50), the equality $w(z_1) = w(z_2)$ gives

$$\frac{az_1 + b}{cz_1 + d} = \frac{az_2 + b}{cz_2 + d}. \tag{25.52}$$

If $cz_1 + d = 0$ the left-hand side is infinite, so the right-hand side must be infinite as well. This requires $cz_2 + d = 0$, so then $z_1 = z_2$. If the denominators do not vanish, we can cross multiply, and, after cancelling common terms, we find:

$$(ad - bc)z_1 = (ad - bc)z_2, \tag{25.53}$$

which, on account of (25.51), implies $z_1 = z_2$. The linear fractional map is an invertible transformation. We can easily solve for z in terms of w to get

$$z = \frac{dw - b}{-cw + a}. \tag{25.54}$$

This is also a linear fractional map, and it is also one-to-one, as you can see by examining the condition in (25.51). The map in (25.50) can be rewritten, after cross multiplication, in the form

$$Awz + Bw + Cz + D = 0, \tag{25.55}$$

for complex constants $A, B, C,$ and D. This is just another form that should be recognized as defining a linear fractional transformation. The symmetric role of w and z is manifest here.

Quick calculation 25.4 What is the condition on the coefficients $A, B, C,$ and D for (25.55) to be a linear fractional transformation?

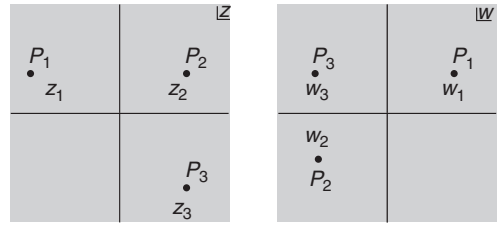

Fig. 25.13 Three punctures P_1, P_2, and P_3 with coordinates z_1, z_2, and z_3, respectively, can be conformally mapped to arbitrary positions w_1, w_2, and w_3 in the w-plane.

How many parameters does the map (25.50) have? Although the map is written using four complex numbers, a, b, c, and d, there are only three parameters. If we multiply a, b, c, and d by the same complex number $\alpha \neq 0$, the map in (25.50) will not change since the constant factor appears both in the numerator and in the denominator. Since $ad - bc \neq 0$, we can rescale the parameters a, b, c, and d by dividing them by $\sqrt{ad - bc}$. The new, rescaled parameters satisfy

$$ad - bc = 1. \tag{25.56}$$

It is now clear that we have only three adjustable complex parameters: given a, b, and c, for example, the equation fixes d.

Having three complex parameters, we might expect that a linear fractional transformation can map a z-sphere with three (different) punctures at z_1, z_2, and z_3, into a w-sphere where the three punctures go to *arbitrary* (different) coordinates w_1, w_2, and w_3 (see Figure 25.13). This is indeed the case, and it shows that $\hat{\mathbb{C}}$ with three punctures is a Riemann surface without moduli.

Let us now construct the map. We write

$$(z - z_1)\,(\cdots) = (w - w_1)\,(\cdots), \tag{25.57}$$

because we want that $w = w_1$ when $z = z_1$. The dots indicate additional multiplicative factors yet to be included. Clearly z_1 will go to w_1 since, regardless of the multiplicative factors, both sides of the equation are equal if $z = z_1$ and $w = w_1$. Next we want $w = w_2$ when $z = z_2$. To achieve this, we put the factors $w - w_2$ and $z - z_2$ in the denominators:

$$\frac{z - z_1}{z - z_2}\,(\cdots) = \frac{w - w_1}{w - w_2}\,(\cdots). \tag{25.58}$$

If $w = w_2$, the equality requires $z = z_2$. In order to incorporate the condition that $z = z_3$ when $w = w_3$, we simply include numerical factors that make each side equal to one when this happens:

$$\frac{z - z_1}{z - z_2}\,\frac{z_3 - z_2}{z_3 - z_1} = \frac{w - w_1}{w - w_2}\,\frac{w_3 - w_2}{w_3 - w_1}. \tag{25.59}$$

The above equation is our final result for the map that takes z_1, z_2, and z_3 to w_1, w_2, and w_3, respectively. It is a linear fractional transformation: by cross multiplying, we find that (25.59) is in the form of (25.55). Spheres with three punctures have no moduli because all

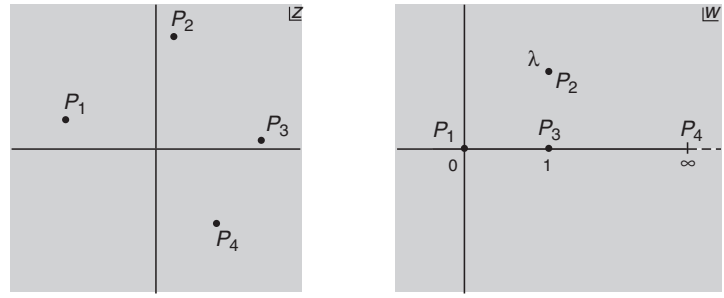

Fig. 25.14 **Four punctures in the complex z-plane can be mapped uniquely into the w-plane by requiring that P_1, P_3, and P_4 go to $w = 0$, 1, and ∞, respectively. In this canonical presentation, the remaining puncture P_2 will go to some value $w = \lambda$.**

such spheres are conformally equivalent; they can be mapped into each other using (25.59). In Problem 25.1 you will show that (25.59) is the *unique* map of $\hat{\mathbb{C}}$ taking an ordered set of three different points to another ordered set of three different points.

Consider now $\hat{\mathbb{C}}$ with four punctures at z_1, z_2, z_3, and z_4, and another copy of $\hat{\mathbb{C}}$ with *four* punctures at w_1, w_2, w_3, and w_4. Think of the punctures as labeled punctures, with labels 1, 2, 3, and 4. Can these copies of $\hat{\mathbb{C}}$ be mapped into each other with the punctures going into each other, that is, with $z_i \to w_i$, for $i = 1, 2, 3$, and 4? In general the answer is no. To map the four punctures into each other, we must certainly map z_1, z_2, and z_3 into w_1, w_2, and w_3. This requirement fixes the map (25.59), giving us the unique map capable of mapping the punctures. Being unique, we cannot incorporate an additional condition, and z_4 will go to some point that we cannot prescribe. Of course, the map may happen to send z_4 to the prescribed w_4, but that would be accidental, since z_4 and w_4 were not used to construct the map. Therefore, spheres with four labeled punctures are not necessarily conformally equivalent.

Suppose someone hands you a z-sphere and a z'-sphere, each with four punctures P_1, P_2, P_3, and P_4. How do you decide whether they are conformally equivalent? To find out, first take each sphere and map P_1, P_3, and P_4 to 0, 1, and ∞. This requires linear fractional maps $w = f(z)$ and $w' = f'(z')$. The question of conformal equivalence of the z- and z'-spheres is now the question of conformal equivalence of the w- and w'-spheres. Now look at P_2. Suppose P_2 is sent to $w = \lambda$ in the w sphere and to $w' = \lambda'$ in the w'-sphere. The four-punctured w- and w'-spheres are conformally equivalent *if and only if* $\lambda = \lambda'$. The only conformal map $w \to w'$ preserving the punctures P_1, P_3, and P_4 (set at 0, 1, and ∞ in both) is the identity map. If the P_2 punctures are to go into each other, their coordinates must be the same. This establishes the claim. In the canonical presentation, where three punctures are sent to 0, 1, and ∞, the position λ of the last puncture is therefore a modulus of the four-punctured sphere (see Figure 25.14). Since this position is a complex number, we have a complex modulus, or equivalently, two real moduli. These moduli are continuous parameters, since the position of the fourth puncture can be varied continuously.

The complex modulus λ which represents the position of a puncture can take all complex values except for 0, 1, and ∞, since in those cases two of the punctures would coincide. Shifting the viewpoint, consider a new Riemann sphere, with coordinate λ, and with the points 0, 1, and ∞ removed. To each point λ_0 of this sphere we associate a sphere with punctures at $0, 1, \infty$ *and* λ_0. With this association, two different points in the λ-sphere represent conformally inequivalent four-punctured spheres. Finally, every four-punctured sphere is associated with a point on the λ-sphere. We call the λ-sphere, with 0, 1, and ∞ removed, the *moduli space* $\mathcal{M}_{0,4}$ of four-punctured spheres. The zero tells us that we are dealing with Riemann spheres (which are surfaces of genus zero), and the four indicates that the surfaces have four punctures, assumed to be labeled. Being itself a two-dimensional surface, the (real) dimensionality of $\mathcal{M}_{0,4}$ is two:

$$\dim(\mathcal{M}_{0,4}) = 2. \tag{25.60}$$

Amusingly, the moduli space of four-punctured Riemann spheres is itself a Riemann surface, a sphere λ with three punctures. Understanding a moduli space means that we have control over all the inequivalent Riemann surfaces of a given type. The moduli space $\mathcal{M}_{0,4}$ will be important for our discussion of closed string amplitudes in Chapter 26. To understand open string scattering, we now consider $\bar{\mathbb{H}}$ with punctures on the real line and the moduli spaces of such Riemann surfaces.

Recall that $\bar{\mathbb{H}}$ is defined as the upper half-plane with the real line included and a "point at infinity" included, as well. This point at infinity is added in the same way as it was added for the Riemann sphere. In $\bar{\mathbb{H}}$, both plus and minus infinity on the real line are the same point, and so is the limit of $r\exp(i\theta)$, for $0 \leq \theta \leq \pi$, when $r \to \infty$. The point at infinity is part of the boundary of $\bar{\mathbb{H}}$; this boundary is a circle. Note that $\bar{\mathbb{H}}$ is conformally equivalent to the unit disk, as established by the map (25.23). In this map, the boundary of $\bar{\mathbb{H}}$ maps onto the boundary of the disk, and the point at infinity is mapped to $\eta = -1$. We can now clearly understand that the infinite strip is conformally equivalent to $\bar{\mathbb{H}}$ with the point $z = 0$ removed, and the point at infinity removed, as well. The point at infinity is removed because it is not attained for any finite time string. The infinite strip is conformally equivalent to $\bar{\mathbb{H}}$ twice punctured at the boundary

Let us consider the self-maps of $\bar{\mathbb{H}}$. A reasonable approach is to investigate the self-maps of $\hat{\mathbb{C}}$ that preserve the upper half-plane $\bar{\mathbb{H}}$. Following (25.50), we consider transformations of the form

$$w = \frac{az + b}{cz + d}. \tag{25.61}$$

If a, b, c, and d are real, the real line in the z-plane maps to the real line in the w-plane. In fact, up to a common phase that cancels in (25.61), the condition that the real line is mapped to the real line implies that (Problem 25.3)

$$a, b, c, d \in \mathbb{R}. \tag{25.62}$$

Finally, there is one additional condition. As before, the map is one-to-one if and only if

$$ad - bc \neq 0. \tag{25.63}$$

Since we are now working with real coefficients, the left-hand side above is a number that is either positive or negative. Since we can only rescale the coefficients using real numbers, we can set $ad - bc = 1$ if the left-hand side is positive, and we can set $ad - bc = -1$ if the left-hand side is negative. It turns out that to preserve $\bar{\mathbb{H}}$ we need the former:

$$ad - bc = 1. \tag{25.64}$$

Indeed, taking the imaginary part of (25.61), we find

$$\Im(w) = \Im\Big(\frac{az+b}{cz+d}\Big) = \frac{1}{2i}\Big(\frac{az+b}{cz+d} - \frac{a\bar{z}+b}{c\bar{z}+d}\Big). \tag{25.65}$$

Simplifying this expression, we get

$$\Im(w) = \frac{1}{2i}\frac{(ad-bc)(z-\bar{z})}{|cz+d|^2} = \frac{ad-bc}{|cz+d|^2}\,\Im(z). \tag{25.66}$$

This equation makes it clear that $\Im(z) \geq 0$ implies $\Im(w) \geq 0$ if and only if $ad - bc > 0$. Therefore, (25.64) guarantees that (25.61) preserves the upper half-plane. All-in-all, the self-maps of $\bar{\mathbb{H}}$ are defined by equations (25.61), (25.62), and (25.64). They are characterized by three real parameters. In $\bar{\mathbb{H}}$ the point at infinity is treated like a regular point; its image under (25.61) is $w = a/c$.

Let us consider a few examples. Is the map

$$w = -z \tag{25.67}$$

a map of $\bar{\mathbb{H}}$ to $\bar{\mathbb{H}}$? It is not: a point z on the positive imaginary axis, for example, will be mapped to a point w on the negative imaginary axis. Therefore (25.67) cannot be written in the form (25.61) with the requisite conditions. Indeed, we can try to write

$$w = \frac{(-1)\cdot z + 0}{0\cdot z + 1} \;\to\; a = -1,\; b = 0,\; c = 0,\; d = 1, \tag{25.68}$$

but this gives $ad - bc = -1$, which violates condition (25.64). How about

$$w = z + 1\,? \tag{25.69}$$

This maps $\bar{\mathbb{H}}$ to $\bar{\mathbb{H}}$. In fact, it simply translates $\bar{\mathbb{H}}$ to the right by one unit. A more interesting map is

$$w = -\frac{1}{z}\,. \tag{25.70}$$

To see why this one works, first note that it can be written in the form (25.61) by choosing

$$w = \frac{0\cdot z + (-1)}{1\cdot z + 0} \;\to\; a = 0,\; b = -1,\; c = 1,\; d = 0. \tag{25.71}$$

Since $ad - bc = 1$, condition (25.64) is satisfied. More directly, write $z = re^{i\theta}$. Then $w = -\frac{1}{z} = -\frac{1}{r}e^{-i\theta}$. What does this transformation look like in the complex plane? Consider a complex number z with positive imaginary part, as shown in Figure 25.15. The vector $\frac{1}{r}e^{-i\theta}$ is obtained from z by reflection across the real axis and scaling. This vector is therefore in the lower half-plane. However, $w = -\frac{1}{z}$, and the negative sign reflects the

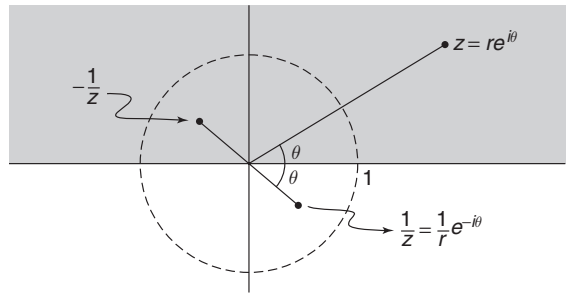

Fig. 25.15 Explaining why the map $z \to -1/z$ takes $\bar{\mathbb{H}}$ onto itself.

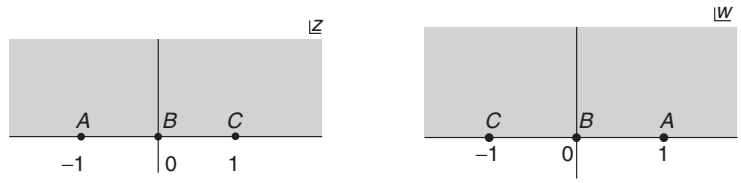

Fig. 25.16 There is no conformal map of the upper half-plane into itself that can reverse the order of three points on the real line.

vector across the origin. The final vector w is therefore in the upper half-plane. It is now clear that (25.70) maps the upper half-plane onto itself.

You might imagine that a real linear fractional transformation, having three real parameters, can be used to map $\bar{\mathbb{H}}$ with three punctures on the real line into $\bar{\mathbb{H}}$ with the three punctures at arbitrary real positions. This is almost true! You can indeed use (25.59) to construct a map, but one thing might go wrong. Condition (25.64) may fail to hold. It is interesting to see why this happens, as it has to do with an important feature of open strings.

The maps (25.61) that we are considering satisfy

$$\frac{dw}{dz} = \frac{(ad - bc)}{(cz + d)^2}, \tag{25.72}$$

and, since $ad - bc > 0$, this derivative is positive throughout the real line. This means that the map takes the *oriented* real line (say from left to right) into a real line with the same orientation! In fact, we can think of the real line in $\bar{\mathbb{H}}$ as a circle that includes the point at infinity. Three points on the real line can be mapped into three other points if and only if their *cyclic* orderings, as points on a circle, agree. Let $(P_1 P_2 P_3)$ denote the punctures 1, 2, and 3, if they appear in this order as we travel on the real line from left to right. Then this configuration can be mapped to any other configuration as long as the cyclic ordering is preserved. It can be mapped to the configuration $(P_2 P_3 P_1)$, for example, but not to the configuration $(P_2 P_1 P_3)$.

Consider an example. Let A, B, and C be points on the z-plane that lie at -1, 0, and 1, respectively (Figure 25.16). Let us try to build a map to the w-plane that takes the points

Fig. 25.17 **The z upper half-plane with four punctures on the real line can be presented canonically by mapping it to the w upper half-plane and setting three of the punctures, say P_1, P_3, and P_4, to $w = 0$, 1, and ∞. The puncture P_2 will have $w = \lambda$ with $0 < \lambda < 1$.**

to $1, 0$, and -1, respectively. In effect, we are trying to reverse the order of the points on the real axis. Using (25.59), we find that we need the following mapping:

$$\frac{z-1}{z+1} = \frac{w+1}{w-1}. \tag{25.73}$$

When we cross multiply, we find that this map is $w = -z$, and we showed before that this does not map $\bar{\mathbb{H}}$ to $\bar{\mathbb{H}}$. Since there are only two inequivalent cyclic orderings of three points on the boundary of a disk, the moduli space $\mathcal{N}_3$ of $\bar{\mathbb{H}}$ with three punctures on its boundary has two representatives. One represents $\bar{\mathbb{H}}$ with three punctures ordered as in $(P_1 P_2 P_3)$, and the other represents $\bar{\mathbb{H}}$ with three punctures ordered as in $(P_2 P_1 P_3)$. We can say that there is one discrete modulus taking two values, and that is all! Note that the moduli space $\mathcal{M}_{0,3}$ of three-punctured spheres had no discrete modulus – all such spheres were conformally equivalent. The moduli space $\mathcal{M}_{0,3}$ is just a point, while the moduli space $\mathcal{N}_3$ is two points. While there is an obvious way to define an ordering for points that lie on a circle (the boundary of $\bar{\mathbb{H}}$), there is no way to define an ordering for points on a sphere.

Consider $\bar{\mathbb{H}}$ with four labeled punctures, P_1, P_2, P_3, and P_4, on the real line, and fix their cyclic ordering to be $(P_1 P_2 P_3 P_4)$ (Figure 25.17). Can we map this $\bar{\mathbb{H}}$ to itself and take the four punctures to arbitrary positions that preserve the cyclic ordering? No. There are just three real parameters in the most general map of $\bar{\mathbb{H}}$ to $\bar{\mathbb{H}}$, and therefore only three positions can be specified. Following the same strategy as before, we can map three of the punctures to fixed w coordinates. For instance, let us map P_1 to $w = 0$, P_3 to $w = 1$, and P_4 to $w = \infty$. Under this map, P_2 will be taken to $w = \lambda$, where λ is some real number. Because the cyclic ordering is preserved, P_2 must still lie between P_1 and P_3, so $0 < \lambda < 1$.

We can view $0 < \lambda < 1$ as the space that represents all possible $\bar{\mathbb{H}}$ with four ordered punctures on the boundary. For any λ_0 in this interval there is an associated $\bar{\mathbb{H}}$ punctured at 0, λ_0, 1 and ∞. Two points with different λ values represent conformally inequivalent $\bar{\mathbb{H}}$s with four boundary punctures. Finally, no $\bar{\mathbb{H}}$ with four boundary punctures is missed in this interval. We thus call $0 < \lambda < 1$ the moduli space $\mathcal{N}_4$ of *upper half-planes with four labeled boundary punctures of a given cyclic ordering*. This moduli space has one real parameter and is just an open interval. It is a one-dimensional space:

$$\dim \mathcal{N}_4 = 1. \tag{25.74}$$

The surfaces $\bar{\mathbb{H}}$ with four boundary punctures are important because they are the surfaces that arise for interactions which involve four open strings. As we will see below, the world-sheet for such a process is conformally equivalent to an $\bar{\mathbb{H}}$ with four boundary punctures. We will also see that the string diagram has one parameter, the time difference T between the interaction points. We will prove a remarkable result: as T varies over its natural range, the string diagrams produce all $\bar{\mathbb{H}}$ with four boundary punctures! That is, we generate the full moduli space $\mathcal{N}_4$. To prove this we must study the function $\lambda(T)$, which gives the modulus as a function of T. This, of course, requires the conformal map from the string diagram to $\bar{\mathbb{H}}$. We must show that, as T goes over its natural range, $\lambda(T)$ ranges from zero to one. We now turn to this analysis.

Quick calculation 25.5 How many inequivalent orderings of four points are possible on the boundary of $\bar{\mathbb{H}}$. What does this tell us about the moduli space of $\bar{\mathbb{H}}$ with four boundary punctures?

25.7 Four open string interaction

Let us consider the process where two open strings join to form an intermediate string, which then splits into two open strings. The incoming strings are strings three and four, and the outgoing strings are strings one and two. The string diagram is shown in Figure 25.18.

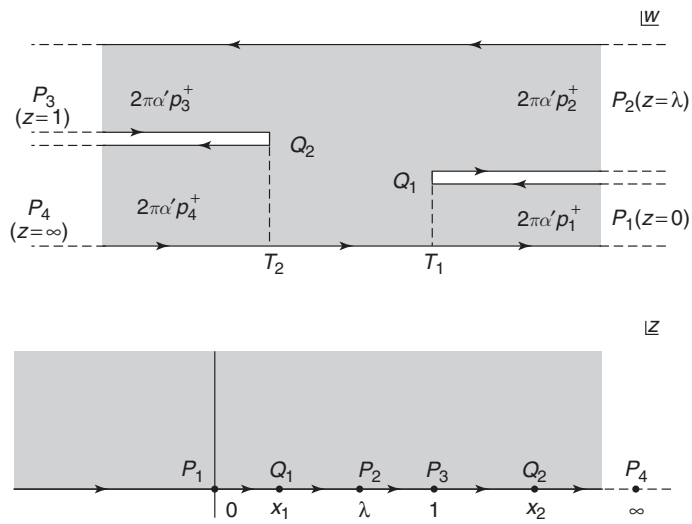

Fig. 25.18 Top: a light-cone string diagram showing strings three and four joining to form an intermediate string that then splits into strings one and two. Bottom: the light-cone diagram is mapped into $\bar{\mathbb{H}}$. P_1, P_3, and P_4 are go to 0, 1, and ∞. The interaction points Q_1 and Q_2 go to x_1 and x_2, and P_2 goes to λ.

Note that the interaction point Q_1 occurs at $\tau = T_1$, and Q_2 occurs at $\tau = T_2$. The time interval between interactions is

$$T = T_1 - T_2. \tag{25.75}$$

In the diagram shown in the figure $T > 0$. But the set of all possible interactions also includes the cases when $T < 0$. In those cases the slits move across each other. The complete range of T for this interaction is

$$-\infty < T < \infty. \tag{25.76}$$

There is an exception if $p_2^+ = p_3^+$. In this case the slits collide when $T = 0$, and T cannot be negative. You may examine such a situation in Problem 25.7.

We now use a Schwarz–Christoffel map to take this diagram into $\bar{\mathbb{H}}$. The special points on the boundary of the string diagram are ordered as $(P_1 Q_1 P_2 P_3 Q_2 P_4)$. Three points can go to arbitrary positions (consistent with the cyclic ordering), so we choose P_1, P_3, and P_4 to go to 0, 1, and ∞, respectively. Then P_2, whose position plays the role of modulus, must map to some $0 < \lambda < 1$. Since Q_1 lies between P_1 and P_2 and Q_2 lies between P_3 and P_4, their images x_1 and x_2 in $\bar{\mathbb{H}}$ satisfy

$$0 < x_1 < \lambda, \quad 1 < x_2 < \infty. \tag{25.77}$$

Our aim is to calculate the modulus λ of the $\bar{\mathbb{H}}$ with four boundary punctures as a function of the parameters T and p_i^+ of the string diagram:

$$\lambda = \lambda\,(T;\ p_1^+, p_2^+, p_3^+, p_4^+). \tag{25.78}$$

We want to prove a remarkable fact: the set of all possible string diagrams $T \in (-\infty, \infty)$ will give all $\lambda \in (0, 1)$, the full moduli space $\mathcal{N}_4$ of $\bar{\mathbb{H}}$ with four boundary punctures. λ is manifestly a well defined function of T: given T, we can calculate λ. Although we will not prove it, the function $\lambda(T)$ is in fact one-to-one. This shows that two string diagrams with different values of T give different Riemann surfaces. As a result, the set of string diagrams gives each surface in $\mathcal{N}_4$ once.

The differential equation for the conformal map is readily written using (25.27). At the points P_i the turning angles are $+\pi$, and at the interaction points Q_i the turning angles are $-\pi$. We therefore have

$$\frac{dw}{dz} = A\,\frac{1}{z}\,(z - x_1)\,\frac{1}{(z-\lambda)}\,\frac{1}{(z-1)}\,(z - x_2). \tag{25.79}$$

We now need the partial fraction expansion of the right-hand side. We can use a trick valid when the factors in the denominator are all different, as they are in (25.79). The coefficient of $1/z$, for example, is obtained by deleting the $1/z$ factor and evaluating the rest of the right-hand side for $z = 0$. Similarly, to determine the coefficient of $1/(z - \lambda)$, you delete this factor, and evaluate the rest at $z = \lambda$. Removing the A factor, the right-hand side is therefore

$$\frac{1}{z}\frac{(-x_1)(-x_2)}{(-\lambda)(-1)} + \frac{1}{(z-\lambda)}\frac{(\lambda - x_1)(\lambda - x_2)}{\lambda(\lambda - 1)} + \frac{1}{(z-1)}\frac{(1 - x_1)(1 - x_2)}{(1 - \lambda)}. \tag{25.80}$$

Simplifying, equation (25.79) becomes

$$\frac{dw}{dz} = A\left[\frac{1}{z}\,\frac{x_1x_2}{\lambda} + \frac{1}{z-\lambda}\,\frac{(\lambda-x_1)(x_2-\lambda)}{\lambda(1-\lambda)} - \frac{1}{z-1}\,\frac{(1-x_1)(x_2-1)}{1-\lambda}\right]. \tag{25.81}$$

We also know a way to write the conformal map directly, as explained below equation (25.46). For the present case it gives

$$w(z) = -2\alpha' p_1^+ \ln z - 2\alpha' p_2^+ \ln(z-\lambda) + 2\alpha' p_3^+ \ln(z-1). \tag{25.82}$$

Taking the derivative,

$$\frac{dw}{dz} = -2\alpha' p_1^+ \frac{1}{z} - 2\alpha' p_2^+ \frac{1}{z-\lambda} + 2\alpha' p_3^+ \frac{1}{z-1}. \tag{25.83}$$

We can compare the two differential equations (25.81) and (25.83) to obtain three equations which relate the string diagram momentum parameters to the parameters in the $\bar{\mathbb{H}}$ presentation:

$$\begin{aligned} -2\alpha' p_1^+ &= A\,\frac{x_1x_2}{\lambda}, \\ -2\alpha' p_2^+ &= A\,\frac{(\lambda-x_1)(x_2-\lambda)}{\lambda(1-\lambda)}, \\ -2\alpha' p_3^+ &= A\,\frac{(1-x_1)(x_2-1)}{1-\lambda}. \end{aligned} \tag{25.84}$$

So far we have three equations to solve for our four unknowns (A, λ, x_1, x_2). What piece of information is missing? We have not yet used the parameter $T = T_1 - T_2$ of the string diagram. Notice that T_1 is the real part of w when $z = x_1$, and T_2 is the real part of w when $z = x_2$:

$$T_1 = \Re(w(x_1)), \quad T_2 = \Re(w(x_2)). \tag{25.85}$$

Since $\Re(\ln z) = \ln|z|$, we can use equation (25.82) to write

$$\begin{aligned} T_1 &= -2\alpha' p_1^+ \ln x_1 - 2\alpha' p_2^+ \ln(\lambda - x_1) + 2\alpha' p_3^+ \ln(1-x_1), \\ T_2 &= -2\alpha' p_1^+ \ln x_2 - 2\alpha' p_2^+ \ln(x_2 - \lambda) + 2\alpha' p_3^+ \ln(x_2-1). \end{aligned}$$

Here we have used $0 < x_1 < \lambda < 1 < x_2$. We can now evaluate $T = T_1 - T_2$:

$$\frac{1}{2\alpha'}\,T = p_1^+ \ln\frac{x_2}{x_1} + p_2^+ \ln\left(\frac{x_2-\lambda}{\lambda-x_1}\right) + p_3^+ \ln\left(\frac{1-x_1}{x_2-1}\right). \tag{25.86}$$

This last equation, together with those previous, fixes the values of the unknown parameters. It is difficult to solve this system of equations to obtain explicit values for the parameters, but, luckily, we can prove our claim without doing so. We will merely analyze them to derive the result.

Since A does not appear in (25.86), it is convenient to eliminate it from equations (25.7) by forming ratios:

$$\begin{aligned}\frac{p_2^+}{p_1^+} &= \frac{(\lambda - x_1)(x_2 - \lambda)}{(1-\lambda)x_1 x_2},\\ \frac{p_2^+}{p_3^+} &= \frac{(\lambda - x_1)(x_2 - \lambda)}{(1-x_1)(x_2-1)\lambda}.\end{aligned} \tag{25.87}$$

As T varies, the values of x_1, x_2, and λ will vary continuously. T varies from $-\infty$ to $+\infty$. If we can show that λ approaches zero as $T \to \infty$ and that λ approaches one as $T \to -\infty$, then λ must take all values $0 < \lambda < 1$.

We have just suggested that $\lambda \to 0$ as $T \to \infty$. The intuition is that, when $T \to \infty$, the points P_1 and P_2 are separated from the points P_3 and P_4 by a long strip. Conformally, P_1 and P_2 are approaching each other. In our map to $\bar{\mathbb{H}}$ this requires $\lambda \to 0$ (if you want to develop your intuition further, solve Problem 25.5). We now prove that $T \to \infty$ as $\lambda \to 0$. Since $\lambda(T)$ is a well defined function of T, this means that $\lambda \to 0$ as $T \to \infty$.

Given that $0 < x_1 < \lambda$, we have $x_1 \to 0$ when $\lambda \to 0$. Let us see what happens to T in this case. Using (25.86), we find

$$\frac{1}{2\alpha'}\, T \simeq\ p_1^+ \ln\frac{x_2}{x_1} + p_2^+ \ln\frac{x_2}{\lambda - x_1} + p_3^+ \ln\frac{1}{x_2 - 1}. \tag{25.88}$$

To understand how T behaves, we must know how x_2 behaves. We use equations (25.7) to find out. With $\lambda \to 0$ and $x_1 \to 0$ they give

$$\begin{aligned}\frac{p_2^+}{p_1^+} &\simeq \frac{(\lambda - x_1)}{x_1} = \frac{\lambda}{x_1} - 1,\\ \frac{p_2^+}{p_3^+} &\simeq \frac{(\lambda - x_1)x_2}{(x_2-1)\lambda} = \frac{1 - \frac{x_1}{\lambda}}{1 - \frac{1}{x_2}}.\end{aligned} \tag{25.89}$$

Since the ratios of light-cone momenta are fixed positive numbers, the top equation shows that, in the limit, λ/x_1 is finite and greater than one. The bottom equation then shows that x_2 is finite and greater than one. Using the finiteness of x_2, (25.88) becomes

$$\frac{1}{2\alpha'}\, T \simeq -p_1^+ \ln x_1 - p_2^+ \ln(\lambda - x_1) + \text{finite}. \tag{25.90}$$

Since both x_1 and $\lambda - x_1$ go to zero as $\lambda \to 0$, we find $T \to +\infty$. Therefore $\lambda \to 0$ when $T \to \infty$, as we wanted to show.

Now we would like to show that $\lambda \to 1$ as $T \to -\infty$. This is also reasonable from the viewpoint of the string diagram. In this case the slits have gone long past each other, so points P_2 and P_3 are separated from points P_1 and P_4 by a long strip. Conformally, P_2 and P_3 are approaching each other. Since P_3 sits at one, this requires $\lambda \to 1$.

Let us look at equation (25.7) to see what is happening with the parameters. As $\lambda \to 1$, the second ratio becomes

$$\frac{p_2^+}{p_3^+} = \frac{(\lambda - x_1)(x_2 - \lambda)}{(1-x_1)(x_2-1)}. \tag{25.91}$$

If we naively take $\lambda \to 1$ in the above equation, we find $p_2^+/p_3^+ \to 1$, which is not correct. Therefore, as $\lambda \to 1$ we must have either $x_1 \to 1$ with x_2 finite or $x_2 \to 1$ with x_1 finite. If $x_1 \to 1$, we get

$$x_1 \to 1: \quad \frac{p_2^+}{p_3^+} \simeq \frac{(\lambda - x_1)}{(1 - x_1)} < 1, \tag{25.92}$$

where the last inequality follows because $x_1 < \lambda$. Having a look at Figure 25.18, we see that we chose conventionally $p_2^+ > p_3^+$. Therefore, we look at the possibility $x_2 \to 1$. In this case, we get

$$x_2 \to 1: \quad \frac{p_2^+}{p_3^+} = \frac{(x_2 - \lambda)}{(x_2 - 1)} > 1. \tag{25.93}$$

This is the correct regime:

$$p_2^+ > p_3^+, \quad \lambda \to 1, \quad x_2 \to 1, \quad \text{and} \quad x_1 \text{ finite}. \tag{25.94}$$

Now we can look at (25.86) to find what T does in this regime:

$$\frac{1}{2\alpha'} T \longrightarrow p_1^+ \ln \frac{1}{x_1} + p_2^+ \ln\left(\frac{x_2 - \lambda}{1 - x_1}\right) + p_3^+ \ln\left(\frac{1 - x_1}{x_2 - 1}\right). \tag{25.95}$$

Separating finite parts, we have

$$\frac{1}{2\alpha'} T \longrightarrow p_2^+ \ln(x_2 - \lambda) - p_3^+ \ln(x_2 - 1) + \text{finite}. \tag{25.96}$$

The first two terms on the above right-hand side can be rearranged to read

$$\frac{1}{2\alpha'} T \longrightarrow (p_2^+ - p_3^+) \ln(x_2 - \lambda) + p_3^+ \ln\left(\frac{x_2 - \lambda}{x_2 - 1}\right) + \text{finite}. \tag{25.97}$$

The ratio appearing on the right-hand side is finite (see (25.93)); therefore

$$\frac{1}{2\alpha'} T \longrightarrow (p_2^+ - p_3^+) \ln(x_2 - \lambda) + \text{finite}\,. \tag{25.98}$$

With $p_2^+ - p_3^+ > 0$, and $x_2 - \lambda \to 0$, we find $T \to -\infty$. Since $T \to -\infty$ as $\lambda \to 1$, we have that $\lambda \to 1$ as $T \to -\infty$, as we wanted to prove.

Since λ varies continuously and reaches both zero and one, it must attain all values in between. We have thus shown that the light-cone diagrams produce all relevant Riemann surfaces, that is, all $\bar{\mathbb{H}}$ with four boundary punctures. The string diagrams have produced the moduli space $\mathcal{N}_4$. As we mentioned before, λ varies monotonically with T, so each value of λ in the interval $(0, 1)$ is taken only once.

The above is an illustration of an important and general result in string theory. For arbitrary string interactions an analogous result holds. Keeping the topological type of the string diagram fixed, as the parameters of the diagram are varied over the natural ranges, the complete set of inequivalent Riemann surfaces of the given topological type is produced. String interactions generate the moduli spaces of Riemann surfaces.

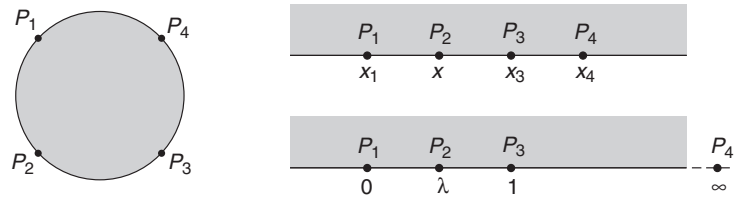

Fig. 25.19 The scattering amplitude for four open string tachyons uses a disk with four boundary punctures. Mapped to $\bar{\mathbb{H}}$, these four punctures P_i, $i = 1,\ldots,4$ have general positions x_1, x, x_3, x_4. They can also be mapped to 0, λ,1, and ∞.

25.8 Veneziano amplitude

We will now use our understanding of the moduli space $\mathcal{N}_4$ to motivate the construction of the Veneziano amplitude, the amplitude for the scattering of four open string tachyons (two incoming and two outgoing). As you recall, open string tachyons are particles with $p^2 = -M^2 = 1/\alpha'$. In Section 25.2 we discussed the three steps necessary to get cross sections in string theory. We have done step (1), the drawing of string diagrams and their conformal mapping into a standard presentation. We now focus on step (2), where we must get the scattering amplitude. We will not consider step (3), which is the construction of a cross section.

We cannot derive here the scattering amplitude, but we can motivate it. For this, we note that it should be an integral over the *moduli* space of the surfaces that contribute to the process – the moduli space $\mathcal{N}_4$. Note that the integral is over the moduli space of surfaces, not over the surfaces themselves! Each surface represents a possible "path" connecting the initial and final states, so each surface contributes to the amplitude. To find the full amplitude, we must add up the contributions from all the surfaces; this is an integral over the parameter space of the surfaces.

Consider $\bar{\mathbb{H}}$ punctured at x_1, x, x_3, and x_4 with $x_1 < x < x_3 < x_4$, as shown in Figure 25.19. These are the points where the four tachyons are supposed to be inserted: the images of the far past and far future strips in the light-cone diagram. At x_1 we introduce the tachyon with spacetime momentum p_1, at x the tachyon with momentum p_2, and at x_3 and x_4 the tachyons with momenta p_3 and p_4, respectively. We could map x_1, x_3, and x_4 to 0, 1, and ∞, but we will not do so yet. Treating all punctures on the same footing makes for a clearer structure. Our strategy will be to use the conditions of conformal invariance to motivate an expression for the amplitude.

Using surfaces with punctures at x_1, x, x_3, and x_4, the moduli space $\mathcal{N}_4$ can be described as the set of surfaces obtained when x varies over the interval $x_1 < x < x_3$ (if you map x_1, x_3, and x_4 to 0, 1, and ∞, you recover the canonical presentation $x \in (0, 1)$). Writing the scattering amplitude $A(p_1, p_2, p_3, p_4)$ as an integral

$$A(p_1, p_2, p_3, p_4) = g_o^2 \int d\mu, \tag{25.99}$$

the measure $d\mu$ must include a dx since we will integrate over x. The amplitude is proportional to the square of the open string coupling g_o, since the process involves two elementary interactions where two strings combine to form one string, and each such interaction carries one power of g_o. This amplitude A is an object similar to the amplitude T considered in (25.2). We will not attempt to fix the precise overall normalization of A. The measure $d\mu$ must also depend on the momenta of the tachyons and on the positions of the boundary punctures. Our momenta satisfy

$$p_1 + p_2 + p_3 + p_4 = 0, \quad p_1^2 = p_2^2 = p_3^2 = p_4^2 = \frac{1}{\alpha'}. \tag{25.100}$$

The first equation is the momentum conservation equation, with all momenta treated as incoming: if the outgoing tachyons are tachyons three and four, for example, then p_3 and p_4 above are actually *minus* the momenta of these two outgoing tachyons. The second equation implies that the four tachyons are on the mass-shell. We now propose a measure $d\mu$:

$$\begin{aligned} d\mu = & dx\, |x_3 - x_1||x_4 - x_1||x_4 - x_3| \\ & |x \ - x_1|^{2\alpha' p_2\cdot p_1} |x_3 - x_1|^{2\alpha' p_3\cdot p_1} |x_3 - x \ |^{2\alpha' p_3\cdot p_2} \\ & |x_4 - x_1|^{2\alpha' p_4\cdot p_1} |x_4 - x \ |^{2\alpha' p_4\cdot p_2} |x_4 - x_3|^{2\alpha' p_4\cdot p_3}. \end{aligned} \tag{25.101}$$

This is a very symmetric expression. For each pair of punctures, there is a factor which is the distance between the points raised to the dot product of their full momenta. In this part of $d\mu$, occupying the second and third lines, all punctures are treated on the same footing. In the first line, however, the moving puncture x is treated differently from the other three fixed punctures, x_1, x_3, and x_4.

Now we consider the issue of conformal invariance. We have written an expression for $d\mu$ that arbitrarily selected three special punctures. We are integrating over the moduli space $\mathcal{N}_4$, presented as $x \in (x_1, x_3)$ for an arbitrary choice of x_1, x_3, and x_4. This choice must be irrelevant – the integral of $d\mu$ must give the same answer for any other choice. Since any two choices of three fixed punctures on the real line are related by a real linear fractional transformation, we can guarantee that the amplitude is unchanged *if the measure $d\mu$ is invariant under real linear fractional transformations*. The generic transformation of this type is

$$z \longrightarrow \frac{az + b}{cz + d}, \quad ad - bc = 1, \quad a, b, c, d \in \mathbb{R}. \tag{25.102}$$

It changes all of the quantities dx, x_1, x_2, x_3. Indeed, we will have

$$\begin{aligned} x &\longrightarrow \frac{ax + b}{cx + d}, \\ dx &\longrightarrow \frac{dx}{(cx + d)^2}, \\ x_i - x_j &\longrightarrow \frac{x_i - x_j}{(cx_i + d)(cx_j + d)}. \end{aligned} \tag{25.103}$$

The measure $d\mu$ must be invariant under all of these changes. This is a stringent condition. Now we can understand the logic behind the first line in (25.8). In including the extra factors along with dx, the first line transforms in a rather neat way. We have, with $x_2 \equiv x$,

$$dx\,|x_3 - x_1||x_4 - x_1||x_4 - x_3| \longrightarrow dx\,|x_3 - x_1||x_4 - x_1||x_4 - x_3| \prod_{i=1}^{4} (cx_i + d)^{-2}. \tag{25.104}$$

The transformation induced exactly the same factor for each of the punctures. Using the last equation in (25.8) to transform the second and third lines in the measure, we find that $d\mu$ transforms into itself times factors of $(cx_i + d)$ for each of the punctures. For $d\mu$ to be invariant, each factor must have the value one. Let us have a look at the factor $(cx_1 + d)$. We get a contribution from the first line of the measure, as calculated in (25.8), and three additional contributions, from the rest of the measure:

$$(cx_1 + d)^{-(2+2\alpha' p_2\cdot p_1 + 2\alpha' p_3\cdot p_1 + 2\alpha' p_4\cdot p_1)}. \tag{25.105}$$

This factor can only be equal to one if the exponent is equal to zero. Happily, this is the case. Looking at the exponent, we have

$$2 + 2\alpha' p_1 \cdot (p_2 + p_3 + p_4) = 2 - 2\alpha' p_1^2 = 2 - 2\alpha' \frac{1}{\alpha'} = 0. \tag{25.106}$$

In obtaining this result, we used both momentum conservation and the on-shell condition (see (25.100)). Completely analogous results hold for the other punctures. This concludes our verification of the conformal invariance of the measure $d\mu$. The condition of conformal invariance is not strong enough to determine the measure uniquely, so we cannot prove that $d\mu$ is selected. It turns out that $d\mu$ is the correct measure for the scattering of open string tachyons.

We simplify the measure by choosing $x_1 = 0$, $x = \lambda$, $x_3 = 1$, and $x_4 = \infty$. It might seem that the placement of x_4 at infinity could give us a problem. If we trust our previous analysis, however, we know that no problem can arise. The measure $d\mu$ is finite for finite values of x_1, x_3, and x_4. Any linear fractional transformation, even if it takes one point to infinity, will not change the measure. Using $x_4 \gg x_1, x, x_2$, we can simplify (25.8):

$$d\mu = d\lambda\,|x_4|^2\,|\lambda|^{2\alpha' p_2\cdot p_1}\,|1-\lambda|^{2\alpha' p_3\cdot p_2}\,|x_4|^{2\alpha' p_4\cdot(p_1+p_2+p_3)}. \tag{25.107}$$

The total exponent of $|x_4|$ adds up to zero, so $d\mu$ reduces to

$$d\mu = d\lambda\,|\lambda|^{2\alpha' p_2\cdot p_1}|1-\lambda|^{2\alpha' p_3\cdot p_2}. \tag{25.108}$$

Since $\lambda \in (0, 1)$, we find that the Veneziano amplitude is given by

$$A(p_1, p_2, p_3, p_4) = g_o^2 \int_0^1 d\lambda\,\lambda^{2\alpha'(p_1\cdot p_2)}(1-\lambda)^{2\alpha'(p_2\cdot p_3)}. \tag{25.109}$$

String theory began with this formula, written by Veneziano in the late 1960s. The formula was simply postulated, and physicists wondered what kind of theory would give rise to such an amplitude. It took a few years before it was demonstrated that this amplitude arose from a theory of strings.

Towards the end of Section 25.2, we realized that one can determine what kinds of particles exist in a theory by looking for poles in the scattering amplitudes. We shall do this for the Veneziano amplitude, just as physicists did shortly after it was discovered. As a first step, we express the momentum dot products in terms of the Lorentz invariants s and t. If strings one and two join to form an intermediate string, the string will carry momentum $p_1 + p_2$. On the other hand, if strings two and three join to form an intermediate string, the string will carry momentum $p_2 + p_3$. Since the punctures are cyclically ordered, there are no other possibilities: string one cannot join with string three to form an intermediate string (nor can string two join string four). Thus the relevant invariants are

$$s = -(p_1 + p_2)^2, \quad t = -(p_2 + p_3)^2. \tag{25.110}$$

Expanding s, and using the on-shell condition, we find

$$s = -p_1^2 - p_2^2 - 2p_1 \cdot p_2 = -\frac{2}{\alpha'} - 2p_1 \cdot p_2, \tag{25.111}$$

which leads to

$$2\alpha' p_1 \cdot p_2 = -\alpha' s - 2 = -(\alpha' s + 1) - 1. \tag{25.112}$$

For convenience, define

$$\alpha(s) \equiv \alpha' s + 1, \tag{25.113}$$

so that we may express (25.112) as

$$2\alpha' p_1 \cdot p_2 = -\alpha(s) - 1. \tag{25.114}$$

Using this relation, and the corresponding one for t, the Veneziano amplitude (25.109) is written as

$$A(p_1, p_2, p_3, p_4) = g_o^2 \int_0^1 d\lambda\, \lambda^{-\alpha(s)-1}(1-\lambda)^{-\alpha(t)-1}. \tag{25.115}$$

This integral can be expressed in terms of gamma functions. Indeed, since

$$\int_0^1 dx\, x^{a-1}(1-x)^{b-1} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \tag{25.116}$$

we find that

$$A(p_1, p_2, p_3, p_4) = g_o^2\, \frac{\Gamma(-\alpha(s))\,\Gamma(-\alpha(t))}{\Gamma(-\alpha(s) - \alpha(t))}. \tag{25.117}$$

To find the poles of the amplitude A, we must know about the poles and zeros of the gamma function. We had a first look at the gamma function in Section 3.4, where we proved the recursion relation

$$\Gamma(z+1) = z\,\Gamma(z), \quad \Re(z) > 0. \tag{25.118}$$

This equation shows that the gamma function has a pole at $z = 0$. Indeed, $\Gamma(1+\epsilon) = \epsilon\Gamma(\epsilon)$ with ϵ small gives

$$\Gamma(\epsilon) = \frac{1}{\epsilon}\Gamma(1+\epsilon) = \frac{1}{\epsilon}(\Gamma(1) + \mathcal{O}(\epsilon)) = \frac{1}{\epsilon} + \mathcal{O}(1). \tag{25.119}$$

We can extend the definition of the gamma function to negative real numbers using analytic continuation: equation (25.118) is taken to hold even when $\Re(z) \leq 0$. We then find, for example, a pole at $z = -1$:

$$\Gamma(\epsilon) = (-1+\epsilon)\Gamma(-1+\epsilon) \;\rightarrow\; \Gamma(-1+\epsilon) \simeq -\frac{1}{\epsilon}. \tag{25.120}$$

Similar reasoning shows that the gamma function has poles at all of the negative integers (see also Problem 3.6). It is a famous (and nontrivial) result that the gamma function has no zeros.

We are now ready to investigate the pole structure of the Veneziano amplitude (25.117). Having no zeros, the gamma function in the denominator contributes no poles. All poles must arise from the gamma functions in the numerator. We have poles when

$$\alpha(s) = n, \quad n = 0, 1, 2, 3, \ldots, \tag{25.121}$$

and when

$$\alpha(t) = m, \quad m = 0, 1, 2, 3, \ldots. \tag{25.122}$$

Let us focus on the poles in the s channel. Recall that a pole for $s = M^2$ implies a particle of mass-squared equal to M^2 (see (25.8)). Therefore, in order to find which particles are present in the Veneziano amplitude, we need to find the values of s which give the poles in (25.121). Making use of (25.113), this condition is simply

$$\alpha' s + 1 = n \;\rightarrow\; s = \frac{1}{\alpha'}(-1+n) = M_n^2. \tag{25.123}$$

Since n runs from zero to infinity, there is an infinite number of poles, and an infinite number of particles. Note that the values of M_n^2 define precisely the levels of the relativistic string that we have been studying in this book. When $n = 0$ we find the tachyon ($M_0^2 = -1/\alpha'$). When $n = 1$ we find massless particles. For higher n, there are infinitely many massive particles. Exactly the same set of particles appears in the poles that occur in the t channel. It became clear to physicists early on that the model invented by Veneziano was somehow related to relativistic strings. This intuition was confirmed in the early 1970s when, using the light-cone gauge, the string was quantized and its spectrum was elucidated.

Problems

Problem 25.1 Self-mappings of the sphere $\hat{\mathbb{C}}$.

(a) Show that the composition of two linear fractional transformations (25.50) is a linear fractional transformation.

(b) A fixed point of a transformation $w = f(z)$ is a point z_0 such that $f(z_0) = z_0$. Prove that each linear fractional transformation, with the exception of the identity transformation $w = z$, has at most two fixed points.

(c) Show that (25.59) is the unique linear fractional transformation that maps the three different points z_1, z_2, and z_3 into the three different points w_1, w_2, and w_3, respectively. [Hint: let S and T be two distinct linear fractional transformations which satisfy this property, and consider the transformation obtained by composing S with the inverse of T.]

Problem 25.2 Conformal invariant of the sphere $\hat{\mathbb{C}}$ with four punctures.

Consider four points z_1, z_2, z_3, and z_4 in $\hat{\mathbb{C}}$. Write a map $w(z)$ such that $w(z_2) = 0$, $w(z_3) = 1$, and $w(z_4) = \infty$. Define $\lambda = w(z_1)$. Calculate λ, and show that it is given by

$$\lambda = \lambda(z_1, z_2, z_3, z_4) = \frac{(z_1 - z_2)(z_3 - z_4)}{(z_1 - z_4)(z_3 - z_2)}.$$

Verify explicitly that λ is a conformal invariant, that is, $\lambda(\phi(z_1), \phi(z_2), \phi(z_3), \phi(z_4)) = \lambda(z_1, z_2, z_3, z_4)$, where ϕ is any linear fractional transformation.

Problem 25.3 Self-mappings of $\bar{\mathbb{H}}$.

Prove that if a linear fractional transformation $w = (az + b)/(cz + d)$ maps the real line of the z plane into the real line of the w plane, then a, b, c, and d must all be real, except possibly for a common phase factor that can be removed without changing the map $z \to w$.

Problem 25.4 Upper half-plane $\bar{\mathbb{H}}$ with two boundary punctures.

Consider $\bar{\mathbb{H}}$ with two punctures P_1 and P_2 on the real line, with coordinates $z = x_1$ and $z = x_2$, respectively. Consider another copy of $\bar{\mathbb{H}}$ with two punctures P_1 and P_2 on the real line, with coordinates $z' = x_1'$ and $z' = x_2'$, respectively. Are these two surfaces the same Riemann surface? Prove that they are, by exhibiting the conformal map that takes the punctures into each other while preserving $\bar{\mathbb{H}}$. You may have to write two conformal maps, depending on the sign of $(x_2' - x_1')/(x_2 - x_1)$. What is the geometrical significance of this sign?

Problem 25.5 What does it mean for points to be close in a Riemann surface?

The concept of points approaching each other on a Riemann surface is subtle because we can always scale the coordinates. Consider $\bar{\mathbb{H}}$ with boundary punctures. If we only have two punctures, then they cannot be said to be close because we can always map them to arbitrary positions. The same is true for three punctures. For four punctures it is possible to have a notion of punctures coming together. The notion applies to a family of surfaces where punctures move.

Consider $\bar{\mathbb{H}}$ with four punctures P_i, $i = 1, 2, 3, 4$, with real coordinates $x_i(t)$, where $t \in [0, 1]$. This is a family of surfaces parameterized by the parameter t. As t varies, the coordinates of the punctures will vary. Of course, we can make (real) linear fractional transformations which change the values of the $x_i(t)$.

Definition We say that $P_i \to P_j$ as $t \to 0$ $(i \neq j)$ if the coordinate of P_i approaches the coordinate of P_j in a representation of $\bar{\mathbb{H}}$ where P_j and the other two punctures are held fixed.

To understand this definition properly, consider the following example:

$$x_1(t) = 0, \quad x_2(t) = t, \quad x_3(t) = 1, \quad x_4(t) = \infty.$$

It is clear from the definition that $P_2 \to P_1$ as $t \to 0$. Prove that the following statements hold for this family as $t \to 0$.

(a) $P_1 \to P_2$. This means that if one puncture approaches a second one, then the second also approaches the first.
(b) $P_3 \to P_4$. The other two punctures also approach each other!

Prove now, on general grounds, that

(c) The definition proposed above is conformal invariant. That is, if the relevant coordinates approach each other in a given representation of $\bar{\mathbb{H}}$ where P_j and the other two punctures are held fixed, then they will approach each other in any such representation.

Problem 25.6 Closing off the polygon in the Schwarz–Christoffel map.

The differential equation (25.27) does not show the turning angle α_n at P_n because this point has been mapped to $z = \infty$. We aim to understand how this point at infinity works out.

(a) To find the turning angle at $z = \infty$, consider the large-z limit of (25.27):

$$\frac{dw}{dz} \simeq A z^{-\frac{1}{\pi}\sum_{i=1}^{n-1}\alpha_i}.$$

Define $t = -1/z$, and calculate $\frac{dw}{dt}$ as a function of t. Explain why your result shows that the turning angle is α_n.

(b) The differential equation

$$\frac{dw}{dz} = A(z - x_1)^{-\frac{\alpha_1}{\pi}}(z - x_2)^{-\frac{\alpha_2}{\pi}} \cdots (z - x_n)^{-\frac{\alpha_n}{\pi}}, \tag{1}$$

with the last turning point P_n included as a finite point x_n, represents the situation where there is no corner at $z = \infty$. Prove that the polygon closes. For this, show that as we traverse the full real axis x, the change in w is zero:

$$w(x = \infty) - w(x = -\infty) = \int_{x=-\infty}^{x=\infty} dx\, \frac{dw}{dx} = 0\,.$$

[Hints: use (1) and contour deformation. Argue that there is no contribution from half-circles around the x_i and around ∞.]

Problem 25.7 Four open string interaction in a special configuration.

Consider the light-cone open string diagram of Figure 25.18 in the special configuration where all p^+ momenta are equal: $p_1^+ = p_2^+ = p_3^+ = p_4^+$. Let $T = T_1 - T_2$ denote the world-sheet time difference between the interaction points. Calculate the modulus $\lambda(T)$

explicitly. Show that the moduli space $\mathcal{N}_4$ is produced for $T \in [0, \infty)$. Confirm that λ is a monotonic function of T.

Problem 25.8 Proton decay in intersecting brane models.

We aim to show that in intersecting brane models there are no open string diagrams which represent proton decay into leptons and gauge bosons. We focus on the model of Figure 21.7 and restrict ourselves to string diagrams presented as a disk with boundary punctures, where the set of open string states relevant to the process (incoming and outgoing) are inserted. The punctures split the boundary of the disk into components, which are the loci of open string endpoints, and consequently, are labeled by D-branes.

Prove that no consistent assignment of labels to the boundary components is possible given that proton decay involves three quarks (and a number of leptons and gauge bosons), and a quark has one endpoint on a baryonic brane and the other endpoint elsewhere. [Hint: try drawing string diagrams. Your attempts should evolve into a clear and very brief argument.]

26 Loop amplitudes in string theory

To calculate scattering amplitudes with high accuracy, one must include the contribution from diagrams which contain loops that represent virtual processes. In Einstein's theory of gravity these diagrams give rise to ultraviolet divergences, which reveal intractable short-distance phenomena. String theory contains gravity, but there are no such ultraviolet divergences. The Riemann surfaces that are candidates for short-distance problems admit an interpretation where they clearly describe safe, long-distance phenomena. We illustrate this remarkable property for the case of annuli, which are the surfaces relevant to virtual open string processes, and for the case of tori, which are the surfaces relevant to virtual closed string processes.

26.1 Loop diagrams and ultraviolet divergences

When calculating scattering amplitudes in particle physics, one typically uses an approximation scheme in which the strength of the interactions is assumed to be small. The amplitude is then written in terms of a perturbative series expansion in powers of this small interaction parameter. The Feynman diagrams we considered in Chapter 25 and the similar looking string diagrams were all *tree diagrams*. This means that the graphs (see, for example, Figure 25.2) contain no nontrivial closed paths, or loops. Tree diagrams give the first term in the perturbative expansion of scattering amplitudes. To go beyond this lowest-order approximation, one must consider Feynman diagrams with loops.

Consider the Feynman diagram with a loop shown in Figure 26.1. This diagram represents an incoming particle which splits into two particles that rejoin to form an outgoing particle. The two particles with momentary existence are called virtual particles, and their appearance and subsequent disappearance is called a virtual process. When calculating the contribution of such graphs to amplitudes, one typically encounters divergent quantities. These divergences are called *ultraviolet* (UV) divergences if they arise from virtual processes involving very high energies or momenta or, alternatively, very short times or distances. Virtual processes that involve short distances can be roughly represented by graphs with small loops.

Ultraviolet divergences need not be fatal – in many cases they can be dealt with by the processes of regularization and renormalization. But for some theories regularization and renormalization do not work – the most important case being Einstein's theory of gravity. Faced with unrenormalizable ultraviolet divergences, one can sometimes work accurately at low energies using *effective* field theories which, at some loss of predictive power,

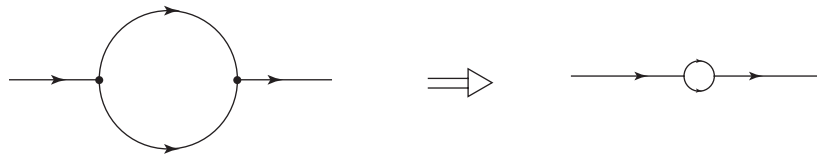

Fig. 26.1 A one-loop Feynman graph that represents a virtual process. In the limit in which the virtual process involves short distances or high momenta, one may find an ultraviolet divergence in the corresponding amplitude.

model in a controllable way the problematic high energy processes. At a fundamental level, however, this is far from being a fully satisfactory approach.

If string theory is a complete quantum theory, then there are two possibilities: either the ultraviolet divergences are manageable or there are simply no ultraviolet divergences. The wonderful news is that *there are no ultraviolet divergences in string theory*. String theory is the first quantum theory, ever, to include gravity without also including ultraviolet divergences. This is the reason why string theory has become the foremost candidate for a theory of quantum gravity. It is the purpose of this chapter to give you some understanding of this remarkable result.

The main property of string diagrams is that the relevant information is all encoded in the *Riemann surfaces* they define. Any conformal map of a string diagram gives an *equivalent* physical picture, since the Riemann surface is not changed. We will use conformal mapping to show that string diagrams that may appear to represent potentially dangerous short-distance physics are actually equivalent to diagrams that are manifestly free of ultraviolet problems. We learned in Chapter 25 that the string amplitude for a given process is obtained as an integral over the relevant moduli space of Riemann surfaces. In this integration, we are adding up the contributions to the amplitude from all the surfaces that are consistent with the initial and final states. The integration is not over the surfaces themselves.

For string loop diagrams, the simplest relevant Riemann surfaces are annuli and tori. We will look at both types of surfaces in some detail, and we will study their corresponding moduli spaces: the space of all possible annuli and the space of all possible tori. We will examine the regions of moduli space where Riemann surfaces seem to involve short-distance physics, and by using suitable conformal maps we will show that they represent safe, long-distance physics. We will take this as evidence that there is no room in string theory for ultraviolet divergences. While we will only examine one-loop string diagrams, this remarkable property of Riemann surfaces and their moduli spaces holds for string diagrams with an arbitrary number of loops.

26.2 Annuli and one-loop open strings

Let us begin by considering the one-loop open string diagram shown in Figure 26.2(a). This diagram involves one incoming open string, one outgoing open string, and two intermediate

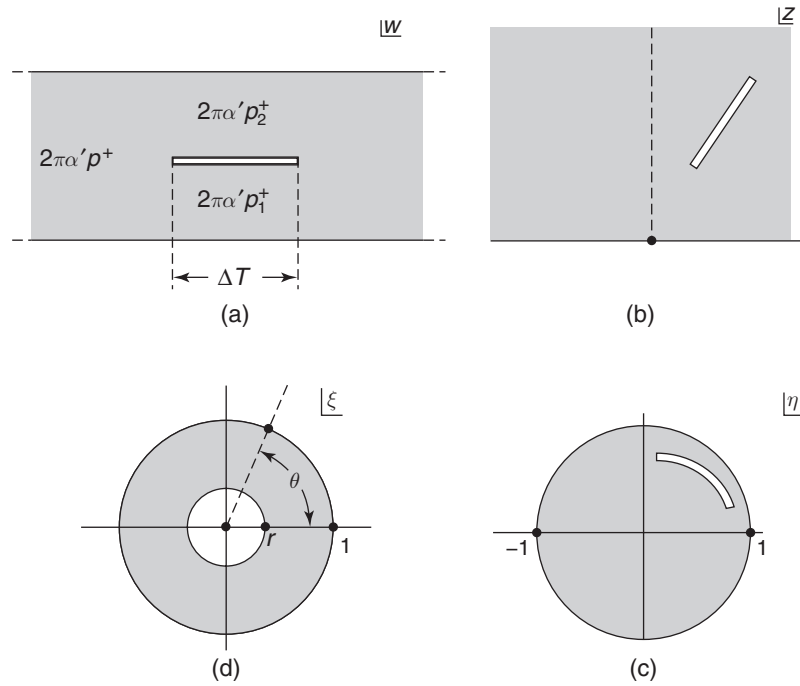

Fig. 26.2 **(a) A string diagram that represents a one-loop open string process. This diagram has two parameters. (b) Mapping the diagram to $\bar{\mathbb{H}}$ with a slit. (c) Mapping the diagram into a disk with a slit. (d) Finally, mapping the diagram into a canonical annulus.**

open strings. It is the string analog of the one-loop Feynman diagram in Figure 26.1. We will discuss a less obvious string analog in Section 26.4.

The light-cone diagram shows an open string with light-cone momentum p^+ that splits into two open strings. The two open strings propagate for a time ΔT and then recombine. By conservation of light-cone momentum we have $p^+ = p_1^+ + p_2^+$. For fixed external momentum p^+, this diagram has two parameters: $\Delta T \in (0, \infty)$, and the vertical position of the slit, parameterized by $p_1^+ \in (0, p^+)$. In the language developed in Chapter 25, we say that the class of Riemann surfaces corresponding to this process has two moduli. The amplitude is obtained by summing the contributions from all such string diagrams.

We now examine a sequence of conformal maps that turns this diagram into a canonical presentation. The first two steps focus on the full strip. As you saw in Section 25.4, the infinite strip can be mapped to $\bar{\mathbb{H}}$ using the exponential map in (25.21). This map, applied in the present case, gives the result indicated in Figure 26.2(b), where the new slit is the image of the original slit under the map. The incoming string has been sent to $z = 0$ and the outgoing one to $z = \infty$. With the help of (25.23) this surface can be mapped into a unit disk – again the slit goes along for the ride, and the incoming and outgoing strings end up at $\eta = 1$, and $\eta = -1$, respectively. At this stage, the string surface is topologically an annulus, a disk with a hole. The hole happens to be a slit.

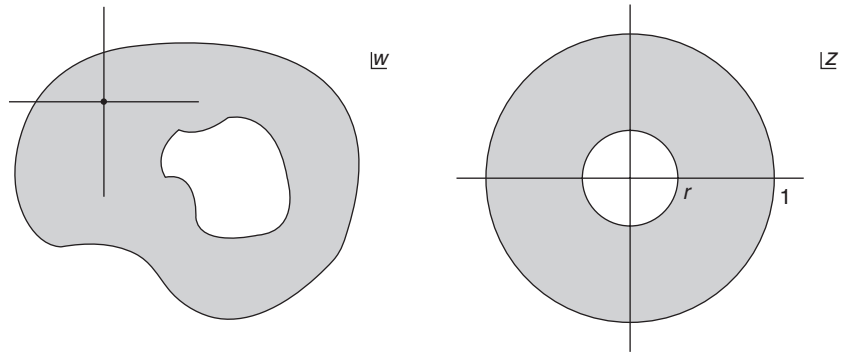

Fig. 26.3 **A topological annulus – the region between the two closed curves in the *w* plane – can be mapped into a canonical annulus $r \leq |z| \leq 1$ for some value of the modulus *r*.**

The last map is fairly nontrivial. The annulus shown in (c) can be mapped conformally into the canonical annulus shown in (d). A canonical annulus with parameter r in the range $0 < r < 1$ is the region $r \leq |\xi| \leq 1$ of the complex ξ plane. This map is a particular example of a general result: a region of the complex plane that is topologically an annulus can be mapped conformally to a canonical annulus (see Figure 26.3). The value r of the inner radius cannot be prescribed; the map fixes it uniquely. This quantity r is called the *modulus of the annulus*. In Figure 26.2(d) the incoming string can be placed at $\xi = 1$: if it originally lies elsewhere on the unit circle, the conformal map $\xi \to \exp(i\alpha)\xi$ will take it to $\xi = 1$ for suitable α. Once the ingoing string has been fixed, the position of the outgoing string cannot be adjusted. In the figure, the outgoing string appears as a puncture at an angle θ with respect to the positive horizontal axis. The Riemann surface is an annulus with two punctures. The boundary of an annulus has two components: the two circles which bound the annular region. In this string diagram, the two punctures lie on the same boundary component.

In the following section, we will use ideas from electrostatics to show explicitly that there exists a conformal map from any topological annulus to a canonical annulus. The modulus of an annulus is closely related to the capacitance of the cylindrical capacitor whose cross sectional region between the conductors is the annulus. We will also learn that two canonical annuli with inner radii r_1 and r_2 cannot be mapped conformally into each other unless $r_1 = r_2$. This is why the inner radius is properly called a modulus. Since the inner radius must be larger than zero and smaller than one, the moduli space of annuli is the space $0 < r < 1$.

Since the string diagram in Figure 26.2(a) has two parameters, the final diagram in part (d) also has two parameters. They are the modulus r of the annulus and the angle θ that defines the position of the outgoing string on the outer boundary. In fact, these two are the moduli of an annulus with two boundary punctures. To construct a string amplitude, we add up the contributions from all the Riemann surfaces that comprise the relevant moduli space. For the Veneziano amplitude we considered all disks with four boundary

punctures. For the open string loop amplitude we are studying now, we must integrate over the moduli r and θ. The integration may be complicated by the appearance of an ultraviolet divergence.

A candidate ultraviolet divergence appears if the string diagrams in some region of the moduli space have a certain kind of short curves that suggest that they may give unbounded contributions to the amplitudes. Let us elaborate.

In the string diagram of Figure 26.2(a), the open strings are vertical segments, some of which stretch between the outer edges of the string diagram, and some of which stretch between an outer edge and the slit. Remember, the slit is one boundary component, and the outer edges of the string diagram, which are separated by the punctures, comprise the other boundary component. All the open strings are open curves that are nontrivial in the following sense: they cannot be contracted away if the endpoints remain on the boundary and are not allowed to slide past the punctures (the external states). Let us call *any* such nontrivial open curve a *possible* open string. There are many more possible open strings than there are open strings. Any curve that begins on the top edge and ends on the lower edge is a possible open string. So is any curve that begins on the left side of the top edge, goes under the slit, and ends on the right side of the top edge. As the name suggests, a possible open string is a curve that could be taken to be an open string with some suitable parameterization of the world-sheet. The concept of a possible open string on a Riemann surface is well defined: the nontrivial character of a possible open string is not changed by conformal mapping. Similarly, a possible closed string is a closed curve that cannot be shrunk away. In Figure 26.2(a), a closed curve looping around the slit is a possible closed string. In this string diagram there are no canonical closed strings, but there are many possible closed strings. The signal of short-distance physics that suggests ultraviolet divergences is the appearance of short possible open strings or short possible closed strings on a string diagram. The length is calculated in the obvious way: in complex coordinates z, a segment dz is assigned length $|dz|$. Since length is not conformal invariant, we can see that the appearance of short possible strings may have an alternative interpretation.

Let us now return to our example and search for the short possible strings that can arise as the modulus r of the annulus varies over its range. Interesting things happen near the ends of the moduli space $0 < r < 1$. As shown in Figure 26.4, an annulus with $r = 1 - \epsilon$ and $\epsilon \to 0$, is a vanishingly thin ribbon. The short radial lines going from one boundary to the other are short possible open strings of length ϵ. The diagram can be interpreted as a short open string traveling around the annulus. But by using a conformal map that scales the figure by the large factor $1/\epsilon$, we get a physically equivalent situation where an open string of unit length travels a very *long* distance $2\pi/\epsilon$. This region of moduli space has therefore the interpretation of long-distance open string physics and is not ultraviolet problematic. The existence of this "dual" interpretation is not the cure for a problem; rather, it simply reveals that the problem is not there. As you may imagine, the region $r \to 1$ corresponds to $\Delta T \to \infty$ in the original light-cone diagram. This is because the slit, which is the inner boundary of the annulus, is becoming very large.

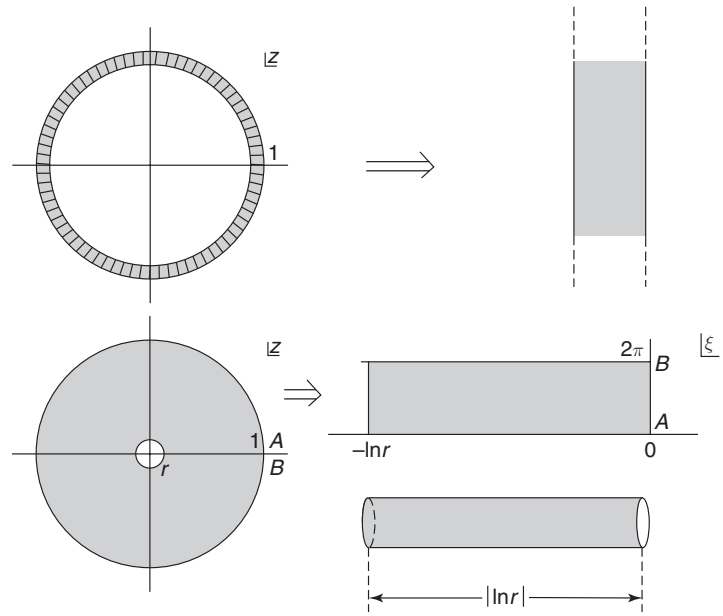

Fig. 26.4 **Top: an annulus with $r \to 1$. To the right we show a small region near $z = 1$ magnified by a large factor to give an open string of unit length that travels a long distance. Bottom: an annulus with $r \to 0$. Its physical interpretation is that of a closed string traveling a long distance.**

The annuli in the region $r \to 0$ are particularly interesting. An $r \simeq 0$ annulus has a very small hole (see the bottom part of Figure 26.4), and we get short possible closed strings that go around the hole. This time we use the map $\xi = \ln z$ to give a nicer physical interpretation. This map turns the annulus into a long strip, with the top and bottom horizontal lines identified. This is simply a long cylinder of length $(-\ln r)$ and circumference 2π. Its physical interpretation is that of a closed string of circumference 2π propagating for a large distance $|\ln r|$. This shows that the region $r \to 0$ of the moduli space of annuli represents *long-distance* closed string physics. The short possible closed curves do not give rise to ultraviolet problems.

These results generalize to arbitrary string diagrams that involve any number of open and closed strings. Every time short possible open strings arise in a light-cone diagram, the diagram can be presented as one where a finite length open string propagates for a long time. Every time short possible closed strings arise in a light-cone diagram, the diagram can be presented as one where a finite length closed string propagates for a long time. A little more is actually required: the presentations in the various regions of the relevant moduli space must go continuously into each other as the moduli vary. This can be done. One can give a presentation that varies continuously with the moduli and that has two properties: (i) open and closed strings that propagate for long distances are represented by strips of constant width and by cylinders of constant circumference, respectively, and (ii) no short possible strings ever arise. This is the geometrical basis for the understanding of the absence of ultraviolet divergences in string theory.

26.3 Annuli and electrostatic capacitance

An annulus in the complex plane $z = x + iy$ is defined as the region between two disjoint closed curves. The curves must have no self-intersections, and one must lie inside the region bounded by the other. We want to use physical ideas from electrostatics to show that such an annulus can be mapped into a canonical annulus, a region $r \leq |\xi| \leq 1$ in the complex ξ plane. The inner radius r is the parameter which defines the annulus. We want to show that r is a modulus for annuli; that is, two annuli with different values of r are not conformally equivalent.

Our electrostatic approach uses the notion of capacitance and some properties of analytic functions, which we review now. As mentioned in Section 25.4, an analytic function $f(z)$ can be broken into real and imaginary parts,

$$f(z) = u(x, y) + i v(x, y)\,, \tag{26.1}$$

that satisfy the Cauchy–Riemann equations

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}\,, \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}\,. \tag{26.2}$$

By taking additional partial derivatives, one readily verifies that the functions u and v satisfy Laplace's equation $\nabla^2 u = \nabla^2 v = 0$. For example,

$$\Big(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\Big)\, v(x, y) = 0\,. \tag{26.3}$$

Additionally, the gradients of u and v are orthogonal vectors of equal magnitude. Indeed, the gradients are defined as

$$\nabla u = \Big(\frac{\partial u}{\partial x}\,, \frac{\partial u}{\partial y}\Big)\,, \quad \nabla v = \Big(\frac{\partial v}{\partial x}\,, \frac{\partial v}{\partial y}\Big)\,, \tag{26.4}$$

and the Cauchy–Riemann equations imply that we can write

$$\nabla u = \Big(\frac{\partial v}{\partial y}\,, -\frac{\partial v}{\partial x}\Big)\,. \tag{26.5}$$

Using this version of ∇u, one readily sees that $\nabla u \cdot \nabla v = 0$ and $|\nabla u| = |\nabla v|$. Moreover, ∇v is obtained from ∇u by a counterclockwise rotation of 90° (the operation $(a, b) \to (-b, a)$ rotates the vector (a, b) by 90° in the counterclockwise direction). Since ∇u is orthogonal to the lines of constant u and ∇v is orthogonal to the lines of constant v, the orthogonality of ∇u and ∇v implies that the lines of constant u and the lines of constant v are orthogonal. These facts are illustrated in Figure 26.5. Given two points P_1 and P_2, the difference in u values $u(P_2) - u(P_1)$ is related to the flux of ∇v across any curve joining the two points:

$$u(P_2) - u(P_1) = \int_{P_1}^{P_2} \nabla v \cdot \vec{n}\, d\ell\,. \tag{26.6}$$

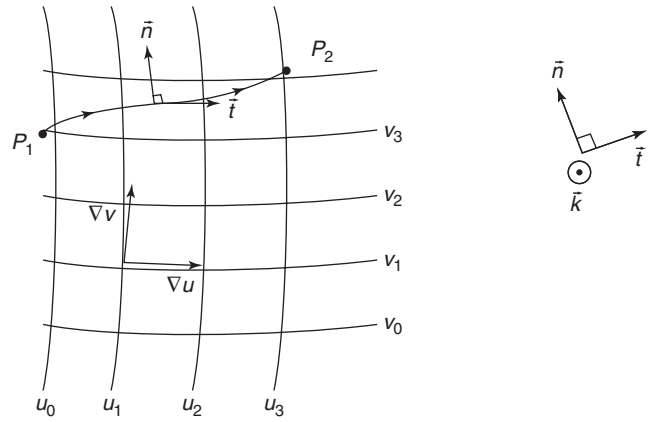

Fig. 26.5 Lines of constant u and lines of constant v. Here $u_0 < u_1 < u_2 < \cdots$ and $v_0 < v_1 < v_2 < \cdots$. The vectors ∇u and ∇v are orthogonal.

Here $d\ell$ is a length element, and $\vec{n}$ is the unit vector normal to the curve obtained by a 90° counterclockwise rotation of the unit tangent vector $\vec{t}$ of the curve, oriented from P_1 to P_2 (see Figure 26.5). The reason is simple: $\nabla v \cdot \vec{n} = \nabla u \cdot \vec{t}$ because the vectors on the left-hand side are obtained by a counterclockwise rotation of 90° from the vectors on the right-hand side, and $\nabla u \cdot \vec{t}$ is the change in u per unit length along the curve.

Having reviewed the basic facts needed, we can now discuss the electrostatic analogy. We will use the topological annulus to define a cylindrical capacitor. The cross section of the capacitor is such that the inner and outer boundaries of the annulus correspond to the inner and outer conductors of the capacitor. The annular region itself is the region between the two conductors. The annulus/capacitor is shown in Figure 26.6 as some region of the complex z plane ($z = x + iy$).

Now imagine setting the inner conductor at unit (electrostatic) potential and the outer conductor at zero potential. With $v(x, y)$ denoting the potential in the region between the conductors, we have the Dirichlet boundary conditions $v = 1$ and $v = 0$ at the inner and outer boundaries, respectively. The electric field in the annular region is given by

$$\vec{E} = -\nabla v\,. \tag{26.7}$$

As you learned in electrostatics, Dirichlet boundary conditions uniquely determine the potential $v(x, y)$ satisfying Laplace's equation in between the conductors. We use physics intuition as sufficient motivation for two facts, which we will not prove: (1) a solution for the potential $v(x, y)$ exists, and (2) the electric field lines begin on the inner conductor and end on the outer conductor.

Since the imaginary part of an analytic function satisfies Laplace's equation, we will identify $v(x, y)$ as the imaginary part of an analytic function $f(z) = u(x, y) + iv(x, y)$ and then construct $u(x, y)$. The lines of constant u will be orthogonal to the lines of constant v, the equipotentials. As a result, the lines of constant u must be the electric field lines.

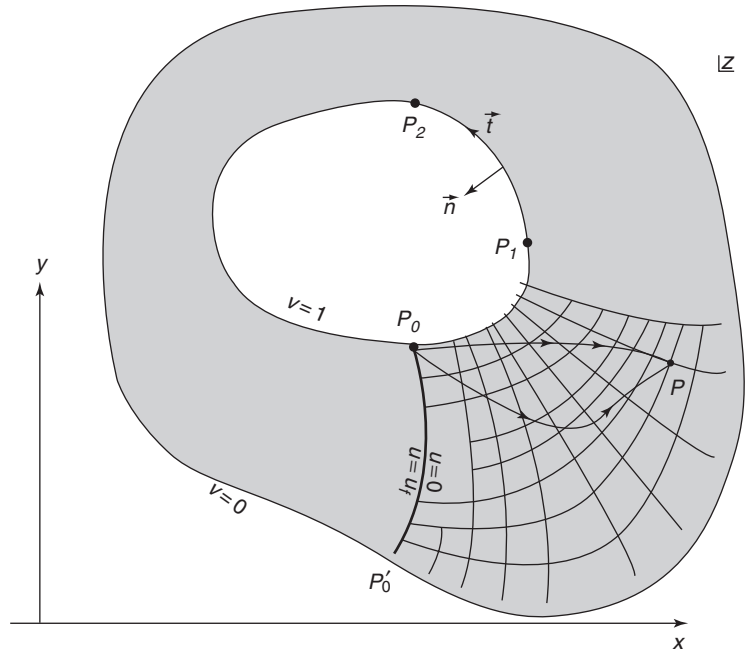

Fig. 26.6 **The cross section of a cylindrical capacitor. The region in between the two conductors defines a topological annulus. We set the inner conductor at unit potential and the outer conductor at zero potential.**

To construct the function $u(x, y)$, we make note of (26.6) and *define*

$$u(P) \equiv \int_{P_0}^{P} \nabla v \cdot \vec{n}\, d\ell\,, \tag{26.8}$$

where P_0 is a fixed point chosen to lie on the inner conductor. The function $u(x, y)$ is well defined in the following sense: if we use two different paths γ_1 and γ_2 from P_0 to P, then the resulting values for $u(P)$ will be the same *if* the two paths can be deformed into each other (see Figure 26.6). In that case the paths γ_1 and γ_2, joined together by reversing the orientation of γ_2, form a closed path $\Gamma = \gamma_1 - \gamma_2$ which bounds a region R. The total integral around Γ then vanishes:

$$\oint_{\gamma_1} \nabla v \cdot \vec{n}\, d\ell - \oint_{\gamma_2} \nabla v \cdot \vec{n}\, d\ell = \oint_{\Gamma} \nabla v \cdot \vec{n}\, d\ell = \int_R da\, \nabla \cdot (\nabla v) = \int_R da\, \nabla^2 v = 0, \tag{26.9}$$

where we have used the divergence theorem to turn the flux of ∇v into the area integral of the divergence of ∇v. This proves that the two paths give the same value for $u(P)$. Note, however, that a path that wraps around the inner conductor need not give the same value for $u(P)$ as a path that does not wrap. This will play a role below.

We now prove that $u + iv$ is an analytic function. For this purpose it is useful to introduce a unit vector $\vec{k}$ which points out of the plane and satisfies $\vec{n} = \vec{k} \times \vec{t}$ (see Figure 26.5). With the help of this vector, we write (26.8) as

$$u(P) = \int_{P_0}^{P} \nabla v \cdot (\vec{k} \times \vec{t})\, d\ell = \int_{P_0}^{P} \nabla v \cdot (\vec{k} \times d\vec{\ell}) = \int_{P_0}^{P} d\vec{\ell} \cdot (\nabla v \times \vec{k})\,, \tag{26.10}$$

where we have used the cyclicity of the triple product. This equation implies

$$\nabla u = \nabla v \times \vec{k}\,. \tag{26.11}$$

The component version of this equation gives precisely the Cauchy–Riemann equations, thus proving the desired result.

Consider now the field line that departs P_0 and reaches the outer boundary at P_0', as shown in Figure 26.6. On any field line, $\nabla v \cdot \vec{n}$ vanishes. It thus follows from (26.8) that the field line in question is a line with $u = 0$. The topology of the annulus, however, implies that u is not single valued. There are paths that cannot be deformed into each other. Using a path that wraps once around the annulus, we find that the field line $u = 0$ is also the line $u = u_f$, with u_f a constant whose value and physical interpretation we will determine next.

The constant u_f can be calculated using definition (26.8) and integrating around the inner conductor, starting at P_0 and moving counterclockwise until we return to P_0:

$$u_f = \oint \nabla v \cdot \vec{n}\, d\ell\,. \tag{26.12}$$

Along the path, $\nabla v = -\vec{E}$ points into the inner conductor, and so does $\vec{n}$. Therefore $\nabla v \cdot \vec{n} > 0$ along the integration curve and, as a consequence, $u_f > 0$. Rewriting the above equation with $\nabla v = -\vec{E}$,

$$u_f = \oint \vec{E} \cdot (-\vec{n})\, d\ell\,. \tag{26.13}$$

Since $(-\vec{n})$ points out of the inner conductor, the above integral computes the flux of $\vec{E}$ out of the inner conductor. By Gauss' law, we find that

$$u_f = Q\,, \tag{26.14}$$

where Q is the charge on the surface of the inner conductor per unit length, where length is measured along the axis of the cylindrical capacitor. The capacitor is neutral, so the outer conductor will have a charge of $(-Q)$ per unit length. For two conductors, with a potential difference V between them and charges Q and $(-Q)$ placed on them, the capacitance C per unit length is defined by the equation $Q = CV$. In our case $V = 1$, and therefore $C = Q$. Using (26.14), we find

$$C = u_f\,. \tag{26.15}$$

The conformal map from the annulus to the canonical presentation is now readily obtained using $v(x, y)$ and $u(x, y)$ to define

$$w = f(z) = u(x, y) + iv(x, y)\,. \tag{26.16}$$

The conformal map $z \to w$ takes the annular region into the configuration of a parallel plate capacitor! This is shown in Figure 26.7. The inner conductor, the surface of constant

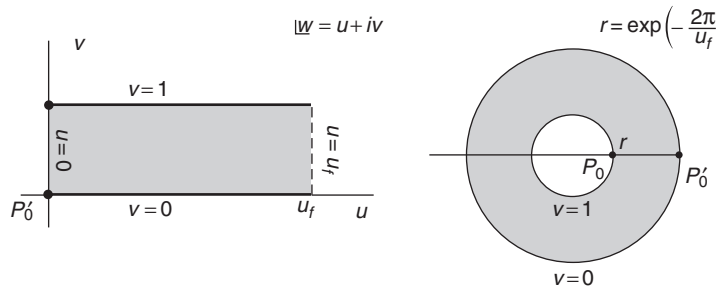

Fig. 26.7 **The capacitor mapped into a parallel plate capacitor and then into a round cylindrical capacitor.**

$v = 1$, is mapped to $\Im(w) = 1$, while the outer conductor, the surface of constant $v = 0$, is mapped to $\Im(w) = 0$. Moreover, the full annulus lies between the vertical lines $u = 0$ and $u = u_f$. The two vertical segments $u = 0$ and $u = u_f$, with $0 \leq v \leq 1$, must be identified, as depicted in Figure 26.6. A final conformal map ξ takes the annular region in the w plane into canonical form:

$$\xi = \exp\left(2\pi i \, \frac{w}{u_f}\right). \tag{26.17}$$

This map takes the outer conductor ($0 \leq u \leq u_f$, $v = 0$) to the unit circle $|\xi| = 1$. It takes the inner conductor ($0 \leq u \leq u_f$, $v = 1$) to the circle $|\xi| = r$, where

$$r = \exp\left(-\frac{2\pi}{u_f}\right) = \exp\left(-\frac{2\pi}{C}\right). \tag{26.18}$$

This relates the inner radius r of the resulting annulus to the constant u_f and the capacitance C. With this final map, we have shown how to use the electrostatic problem to provide a map from the topological annulus to the canonical annulus. The region of the moduli space of annuli interpreted as long-time open string propagation ($r \to 1$) corresponds to large capacitance. The region interpreted as long-time closed string propagation ($r \to 0$) corresponds to small capacitance.

It remains to be shown that r is a modulus. To prove this, we first show that *capacitance is a conformal invariant*. This holds because the electrostatic solution for one capacitor can be used as a solution for any conformally related capacitor. Consider a capacitor in the $z = x + iy$ plane and a capacitor in the $\eta = \psi + i\phi$ plane, as shown in Figure 26.8. Assume that for the capacitor in the z plane we have found an analytic function $f(z) = u + iv$ such that v is the potential when the inner and outer conductors are kept at unit and zero potentials, respectively. Moreover, assume that the conformal map

$$z = h(\eta), \tag{26.19}$$

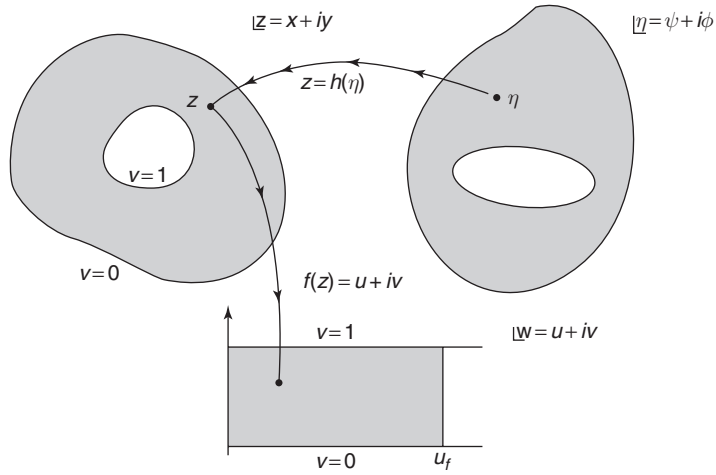

Fig. 26.8 **If the annulus in the η plane can be mapped conformally to the annulus in the z plane, the capacitances of the associated cylindrical capacitors are the same. Here $z = h(\eta)$ denotes the conformal map between the annuli, and $w = f(z)$ maps the z plane capacitor into a parallel-plate capacitor.**

takes the capacitor in the η plane to the capacitor in the z plane. This, of course, means that the inner conductors and the outer conductors are taken into each other by the map. We now claim that the function

$$\begin{aligned} f(h(\eta)) &= u\Big(x(\psi,\phi), y(\psi,\phi)\Big) + i\,v\Big(x(\psi,\phi), y(\psi,\phi)\Big) \\ &\equiv \tilde{u}(\psi,\phi) + i\,\tilde{v}(\psi,\phi)\,, \end{aligned} \tag{26.20}$$

does for the capacitor in the η plane what $f(z)$ did for the capacitor in the z plane. This function assigns to the point η the same number that f assigns to the image of η under the map to the z plane. Since $f(h(\eta))$ is an analytic function, its real and imaginary parts together satisfy the Cauchy–Riemann equations and separately satisfy Laplace's equation. Additionally, the function $\tilde{v}(\psi,\phi)$ will assign unit potential to the inner conductor of the η capacitor and zero potential to the outer conductor. $\tilde{v}(\psi,\phi)$ is the unique solution for the potential in the η capacitor. Up to an additive constant, $\tilde{u}(\psi,\phi)$ is the unique function which can be combined with $\tilde{v}$ to form an analytic function. But the ambiguous constant is unimportant, since the capacitance is a function only of the total change in $\tilde{u}$ as we go around the inner conductor. It is clear that this change is the same as the change in u as we go around the inner conductor of the z capacitor. Thus the capacitance is a conformal invariant.

The conformal invariance of the capacitance implies that the inner radius of a canonical annulus is a modulus. Why? It is clear from (26.18) that two canonical annuli with different inner radii will have different capacitances. Since the capacitance is a conformal invariant, two canonical annuli with different radii cannot be mapped conformally into each other. This is what we wanted to show.

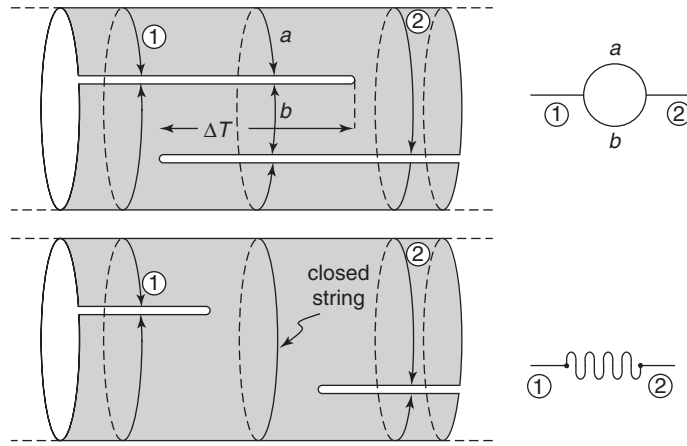

Fig. 26.9 Top: an incoming open string splitting into two strings, string a and string b, that then merge to form string two. The string diagram is nonplanar. Bottom: when the slits do not overlap, we find intermediate closed strings.

26.4 Non-planar open string diagrams

When we studied the one-loop open string diagram of Figure 26.2, we mentioned that there is another string diagram that corresponds to the Feynman diagram of Figure 26.1. That other string diagram is shown in the top part of Figure 26.9. The diagram may appear to be drawn in an unusual way: were it not for the slits, one extending far to the left and the other extending far to the right, we would have a closed string diagram. Both the incoming and the outgoing strings, called string one and string two, are open strings. When the two slits overlap we have two intermediate open strings, string a and string b. The process is that of an open string splitting into two open strings which, later on, recombine to form a single open string.

What is the difference between the one-loop diagram at the top of Figure 26.9 and the one-loop diagram of Figure 26.2? The difference is topological and very significant. In both diagrams, the boundary has two disjoint components. In Figure 26.2, both external open strings are attached to the same boundary component, and the other boundary component is provided by the intermediate slit. This is manifest in Figure 26.2(d). In Figure 26.9 the situation is completely different. The slits themselves are the boundary components, and each boundary component contains an external string. This string diagram is said to be *non-planar* because it cannot be flattened out without tearing it up or having one piece of the surface lie on top of some other piece of the surface. Another non-planar diagram is the infinite cylinder which represents free closed string propagation.

As we learned before, the set of string diagrams that contribute to an amplitude is obtained by varying the parameters that define the string diagram. The diagram at the top of Figure 26.9 has two parameters. The first specifies the light-cone momentum of one of the intermediate open strings. The second, more important for our present considerations,

gives the length ΔT of the time interval during which the slits overlap. When ΔT is large, the slits go past each other for a long time. As $\Delta T \to 0$, the overlap time of the slits goes to zero. But $\Delta T = 0$ is not the end of the parameter space; ΔT can turn negative, in which case we get the string diagram shown in the bottom of Figure 26.9. Surprisingly, we get an intermediate closed string! The string interpretation of the string diagram changes completely. Now the process is described as an open string closing to form a closed string that propagates for some time and then breaks up to form an outgoing open string. To the right of the string diagram we see the corresponding graph. The straight line, representing the open string, turns into the wavy line, representing the closed string. The wavy line then turns back into a straight line. The graph has no loop!

There is an important lesson here. In a consistent theory of interacting quantum open strings, the inclusion of closed strings is generically required. This should not be too surprising. The process by which two open strings join at their endpoints is a *local* process; it is only dependent upon the regions of the strings near the endpoints. This means that it could just as well happen to the endpoints of a single string. Thus, if our theory includes processes in which two strings can join, then it will also, in general, contain processes in which a single open string closes to form a closed string.

A final comment on moduli. When ΔT is large and negative, the two boundaries of the annulus are far from each other. This is the region of moduli space where the inner radius of the canonical annulus is going to zero. This region is dominated by long-time closed string propagation, just as it was in our analysis at the end of Section 26.2. In the former case, the closed string propagates for a long distance and then stops. In the present case, after propagating for a long distance, the closed string turns again into an open string.

26.5 Four closed string interactions

We have seen before that the string diagram for a freely propagating closed string is a twice-punctured Riemann sphere $\hat{\mathbb{C}}$ (Section 25.6). The string diagrams for a closed string interaction which involves two incoming and two outgoing closed strings are Riemann spheres with four punctures. In Section 25.7 we saw that the string diagrams for the interaction of four open strings generate the moduli space of $\bar{\mathbb{H}}$ with four boundary punctures. It is reasonable to expect that the string diagrams for the interaction of four closed strings generate the moduli space $\mathcal{M}_{0,4}$ of four-punctured Riemann spheres. Recall from equation (25.60) that this moduli space has two real parameters. The moduli space is represented by the position of a puncture on $\hat{\mathbb{C}}$ when the other three punctures are fixed, conventionally, to be at 0, 1, and ∞. Our goal in this section is to give some evidence that the hypothesis is, in fact, true: the string diagrams generate the moduli space $\mathcal{M}_{0,4}$. Our arguments will only be qualitative.

Consider the interaction of four closed strings shown in the light-cone diagram of Figure 26.10. Note how closed strings interact: the two incoming closed strings meet at a point and form a longer closed string. After some propagation, two points on the intermediate closed string meet, and the closed string splits into the two outgoing closed strings. If

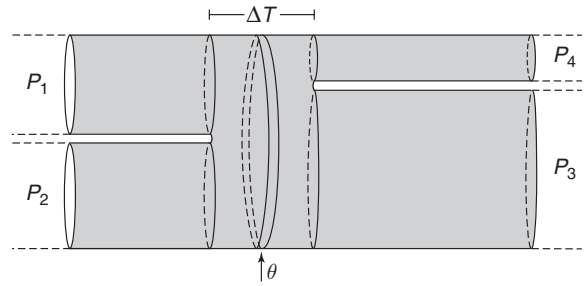

Fig. 26.10 **A light-cone string diagram for the interaction of four closed strings. The Riemann surface is a four-punctured sphere. This string diagram has two parameters, the time ΔT and the twist angle θ. String two approaches string one when ΔT is very large.**

this diagram is to generate even only a portion of the two-dimensional moduli space $\mathcal{M}_{0,4}$, it must have two parameters. One parameter we have encountered already: the time ΔT between the interaction points. For simplicity, this diagram will be taken to include only the range $\Delta T \in [0, \infty)$. We will see that the region $\Delta T < 0$ actually comprises two diagrams! Where is the other parameter of the string diagram? Cut the middle cylinder with a pair of scissors along the double lines shown in the figure. Rigidly rotate the right side of the diagram by an angle θ, and then glue the diagram back together. The resulting diagram does not look like the original one. With a rotation of 90°, for example, string three would be in front of string four. Diagrams obtained from different values of θ can be shown *not* to be conformally equivalent. Therefore $0 \leq \theta < 2\pi$ is the second parameter in the string diagram.

With the parameters $0 \leq \Delta T < \infty$, and $0 \leq \theta < 2\pi$, what region of the moduli space $\mathcal{M}_{0,4}$ does the diagram in Figure 26.10 produce? To answer this question, we imagine mapping this diagram into $\hat{\mathbb{C}}$, letting P_1, P_3, and P_4 land at 0, 1, and ∞, respectively. The modular parameter is the position λ of P_2. How does λ vary as we change ΔT and θ? As $\Delta T \to \infty$, λ will approach P_1, which is at the origin. Moreover, for each fixed ΔT, as θ goes from zero to 2π, λ traces a closed curve around P_1. The string diagram generates a region of $\mathcal{M}_{0,4}$ with the topology of a disk. When $\Delta T = 0$, the curve that we get is the boundary of the disk domain generated by this string diagram.

In the string diagram of Figure 26.10 there is a single intermediate string. There are two diagrams in which the interaction points cross each other and there are three intermediate strings. In the first, shown in Figure 26.11, string one joins string four. In the second, shown in Figure 26.12, string two joins string four.

In Figure 26.11, the time between interactions is $\Delta\tilde{T}$. As $\Delta\tilde{T}$ becomes large, P_2 approaches P_3, which is located at $\lambda = 1$. The intermediate closed string with momentum $p_1^+ - p_4^+$ can be cut, rotated, and glued back, just as before, making the rotation angle θ a parameter. This string diagram will generate a disk domain around $\lambda = 1$. The curve arising from $\Delta\tilde{T} = 0$ represents the boundary of the domain generated by this string diagram.

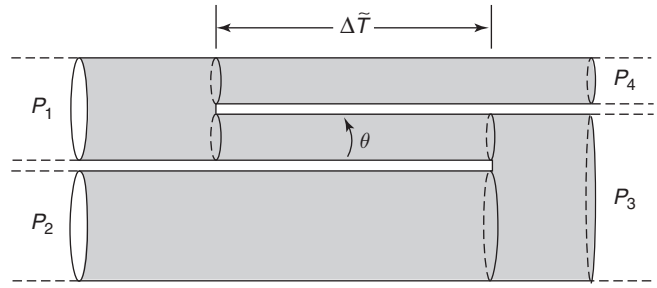

Fig. 26.11 In this string diagram P_2 approaches P_3 when $\Delta\tilde{T}$ becomes very large.

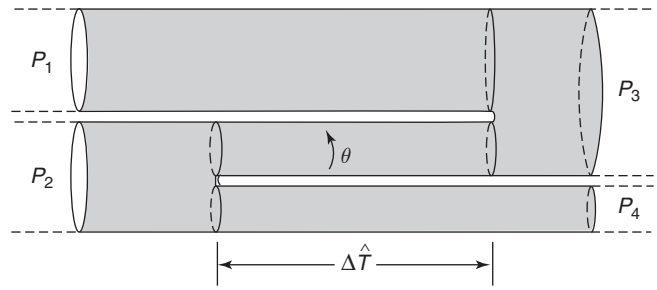

Fig. 26.12 In this string diagram P_2 approaches P_4 when $\Delta\hat{T}$ becomes very large.

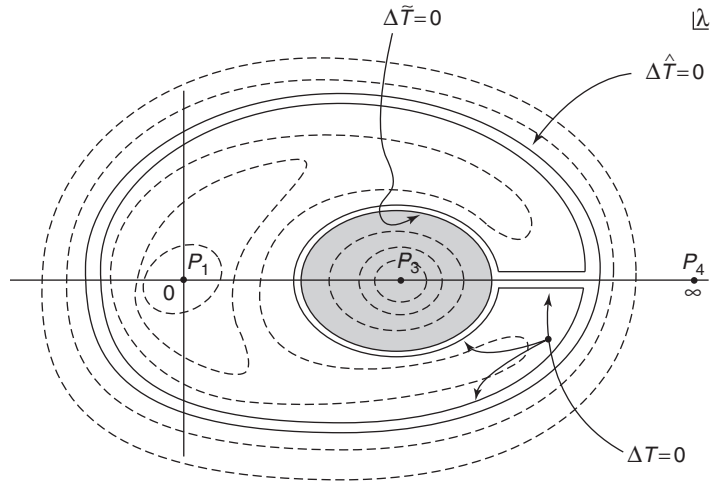

Fig. 26.13 The moduli space $\mathcal{M}_{0,4}$ of four-punctured spheres is generated by three string diagrams. The string diagram in Figure 26.10 generates the disk domain around $\lambda = 0$. The string diagrams in Figures 26.11 and 23.12 generate disk domains around $\lambda = 1$ and $\lambda = \infty$, respectively. The three disk domains fully cover $\mathcal{M}_{0,4}$, which is equivalent to $\hat{\mathbb{C}}$ with three points removed.

In Figure 26.12, the time between interactions is $\Delta\hat{T}$. As $\Delta\hat{T}$ becomes large, P_2 approaches P_4, which is located at $z = \infty$. The intermediate closed string, with momentum $p_2^+ - p_4^+$, also carries a θ parameter. This string diagram will generate a disk domain around $\lambda = \infty$. The curve arising from $\Delta\hat{T} = 0$ represents the boundary of the domain generated by this string diagram.

It turns out that the full moduli space $\mathcal{M}_{0,4}$ is precisely covered by the three string diagrams we have examined, each one producing a punctured-disk domain. Since no region in $\mathcal{M}_{0,4}$ is left unaccounted for, the domain boundaries must meet. So the images in $\mathcal{M}_{0,4}$ of the curves $\Delta T = 0$, $\Delta\tilde{T} = 0$, and $\Delta\hat{T} = 0$ must meet. The way this happens is actually quite intricate, as we show in Figure 26.13. The curves $\Delta\tilde{T} = 0$ and $\Delta\hat{T} = 0$ do not meet at all. Part of the $\Delta T = 0$ curve matches with the curve $\Delta\tilde{T} = 0$, and another part matches with the $\Delta\hat{T} = 0$ curve. Finally, some part of the $\Delta T = 0$ curve matches with itself! The disk domains around $\lambda = 1$ and $\lambda = \infty$ have ordinary shapes, but the one around $\lambda = 0$ has an unusual shape. You may take it as a good challenge to explain the features of Figure 26.13 by looking carefully at the θ dependence of each of the three string diagrams in the limit that the intermediate time goes to zero.

26.6 The moduli space of tori

We conclude this chapter with a look at the fascinating properties of loop amplitudes of closed strings. A light-cone string diagram which represents a one-loop process for a single closed string is shown in Figure 26.14. The slit in this diagram separates two cylinders – these are the intermediate closed strings. If there were no slit, the string diagram would be a two-punctured Riemann sphere. A torus can be viewed as a sphere with a hole. To do so, one can cut out a disk around the north pole of the sphere and a disk around the south pole of a sphere. The two resulting boundaries are then pushed towards each other and glued. The result is a torus. The slit in Figure 26.14 is the hole that turns the sphere into a torus. More precisely, we have a torus with two punctures.

To understand whether ultraviolet divergences can occur, the key questions we must answer are

(1) Why is the torus a Riemann surface?
(2) How many moduli does a torus have?
(3) What does the moduli space of tori look like?

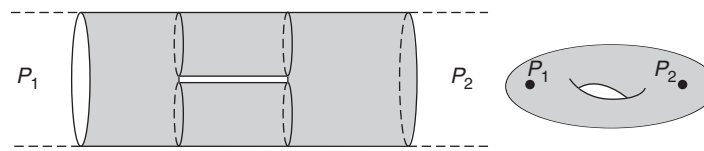

Fig. 26.14 A one-loop light-cone diagram for a closed string. The string diagram is a torus with two punctures.

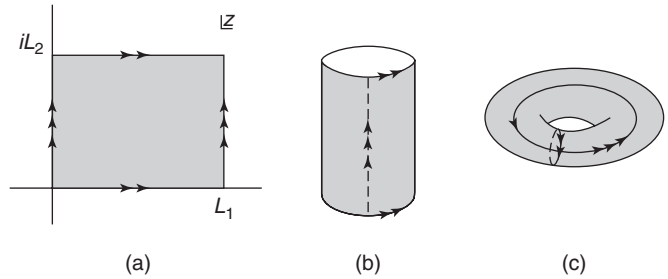

Fig. 26.15 **(a) A rectangular torus is a rectangular region of the complex z-plane with identifications. (b) Gluing the vertical sides of the rectangular region gives a cylinder. (c) Gluing the horizontal sides as well gives a torus.**

Since the answers to these questions are interesting even for tori without punctures, we will consider this case in detail. We will see that tori have two real moduli and that the moduli space is surprisingly intricate.

We begin with the simplest type of torus, the rectangular torus. As shown in Figure 26.15, the torus can be defined as a rectangular region of the complex plane $\mathbb{C}$ with some identifications. In particular, the horizontal sides must be identified as indicated by the double arrows, and the vertical sides must be identified as indicated by the triple arrows. To the right, we show the region after the vertical lines have been identified, and then, in the rightmost part of the figure, we have the final result.

The identifications that lead to the rectangular torus, starting from the full complex plane, are described by the equations

$$z \sim z + L_1\,, \quad z \sim z + iL_2\,. \tag{26.21}$$

The fundamental domain for these identifications is the rectangular region $0 \le \Re(z) < L_1$, $0 \le \Im(z) < L_2$. This torus is a Riemann surface because the complex plane $\mathbb{C}$ is one and because the identifications above are analytic identifications: $z \sim f(z)$, with f an analytic function.

It is important to note that neither L_1 nor L_2 is a parameter of the rectangular torus. We can scale the z coordinate by a constant factor. Letting $z' = z/L_1$, which is obviously a conformal map, the identifications in (26.21) become

$$z' \sim z' + 1\,, \quad z' \sim z' + iT\,, \quad T \equiv \frac{L_2}{L_1}\,. \tag{26.22}$$

It follows that a rectangular torus has just one parameter, the value of T. The above equations define the canonical presentation of a rectangular torus. In the canonical presentation, the horizontal length is one.

Surprisingly, rectangular tori with different T parameters can sometimes be conformally equivalent. To prove this, consider the torus shown in the top left of Figure 26.16. This is a rectangular torus with $T < 1$, shown as a rectangular domain in the w plane. Now we

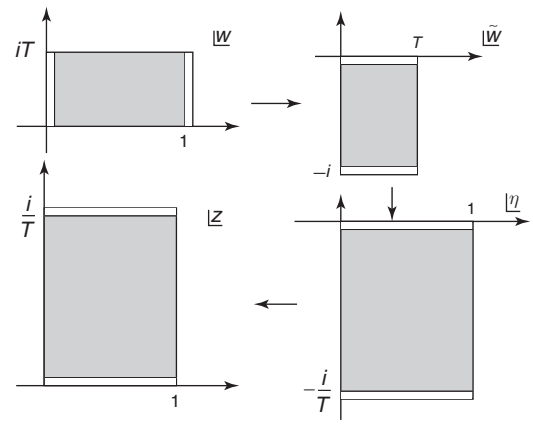

Fig. 26.16 **Top left: a rectangular torus with parameter T in the w plane. Top right: the torus mapped to $\tilde{w} = -iw$. Bottom right: the torus mapped to $\eta = \tilde{w}/T$. Bottom left: the torus mapped to $z = \eta + \frac{i}{T}$. This final form is a rectangular torus with parameter $1/T$.**

perform a sequence of three conformal maps. For visual aid, the vertical lines and their images under the map are shown as double lines. The first map is $\tilde{w} = -iw$. This is a clockwise rotation of the original rectangle by 90°. Next we scale the figure by a factor of $1/T$: $\eta = \tilde{w}/T$. The rectangle now lies below the real line. In the final step, we lift it up: $z = \eta + i/T$. The result (bottom left of the figure) is the canonical presentation of a rectangular torus with parameter $1/T$! We have therefore shown that tori with parameters T and $1/T$ are conformally equivalent. This implies that the tori with $0 < T \leq 1$ are conformally equivalent to the tori with $1 \leq T < \infty$. As a result, the *moduli space* of rectangular tori can be chosen to be the interval $1 \leq T < \infty$ or, alternatively, the interval $0 < T \leq 1$. We are assuming that no conformal map exists that produces a further identification of rectangular tori, which is in fact true.

The above conclusion has some implications for ultraviolet divergences. The tori with $T \to 0$, give rise to short closed strings that appear as vertical segments in the top left of Figure 26.16. On the other hand, these tori are conformally equivalent to large tori with $T \to \infty$, so it is mathematically impossible for them to give rise to ultraviolet problems. In fact, if we include the tori $1 \leq T < \infty$, then we should *not* include the tori $0 < T < 1$. String amplitudes are integrals over the space of inequivalent Riemann surfaces; if we included the short tori we would be double counting. Using the canonical presentation $1 \leq T < \infty$ of the moduli allows us to work with long tori only.

The rectangular tori do not exhaust the set of all conformally inequivalent tori. Tori can be twisted as well. Referring to part (b) of Figure 26.15, one can glue the bottom and top edges of the cylinder with a twist. The way this is actually done is sometimes elusive, so we will explain it in detail. In particular, if you had a real-life rubber cylinder, then gluing its open edges after a twist by an angle θ and after a twist by an angle $\theta + 2\pi$ would be physically inequivalent processes. For a torus Riemann surface, a twist by θ and a twist by $\theta + 2\pi$ are equivalent.

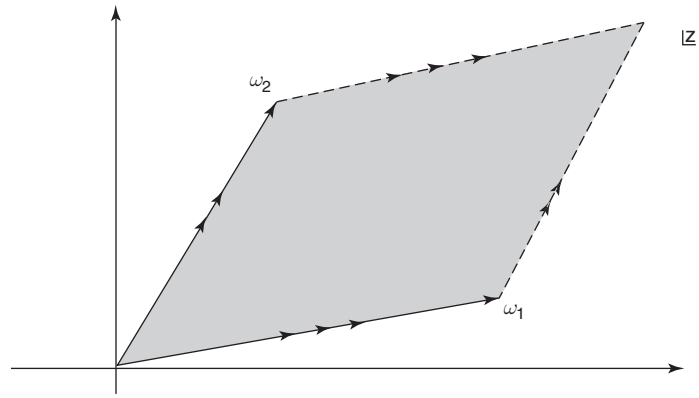

Fig. 26.17 **A general torus is obtained from the complex z-plane by identifying $z \sim z + \omega_1$ and $z \sim z + \omega_2$.**

To discuss twisted tori clearly, we begin by giving the general construction of a torus. To get a torus from the complex plane we need two identifications. We choose two complex numbers ω_1 and ω_2, both different from zero and satisfying

$$\Im\,(\omega_2/\omega_1) > 0\,. \tag{26.23}$$

This condition is simpler than it looks. It just requires that the two complex numbers represent vectors that are not parallel. Then, by convention, we choose ω_2 to be the one which is obtained from ω_1 by a counterclockwise rotation through an angle smaller than 180°. An example is shown in Figure 26.17. The torus is obtained by the identifications

$$z \sim z + \omega_1\,, \quad z \sim z + \omega_2\,. \tag{26.24}$$

The fundamental domain is the parallelogram shaded in the figure, and its edges are identified. This is the general construction of a torus Riemann surface, but it has extraneous parameters. We can define

$$\tau \equiv \omega_2/\omega_1\,, \quad \Im(\tau) > 0\,, \tag{26.25}$$

and, scaling z by a factor of $1/\omega_1$, we note that the identifications above are equivalent to the identifications

$$z \sim z + 1\,, \quad z \sim z + \tau\,, \quad \Im(\tau) > 0\,. \tag{26.26}$$

Observe that τ lives in the upper half-plane $\mathbb{H}$. The new picture of the parallelogram is shown in Figure 26.18. Note that the torus is rectangular precisely when $\Re(\tau) = 0$ (i.e., when τ is purely imaginary).

Could the moduli space of tori be just $\tau \in \mathbb{H}$? It cannot. Our previous work with rectangular tori indicates that tori with different τ parameters can be equivalent. Indeed, for a rectangular torus with parameter T, we have $\tau = iT$, as is clear by comparing (26.22) and (26.26). But this torus is equivalent to a torus with parameter $\tau' = i/T$. It follows that $\tau' = -1/\tau$, and, at least for rectangular tori, τ and $(-1/\tau)$ are conformally equivalent. In fact, we will see shortly that τ and $(-1/\tau)$ always define conformally equivalent tori.

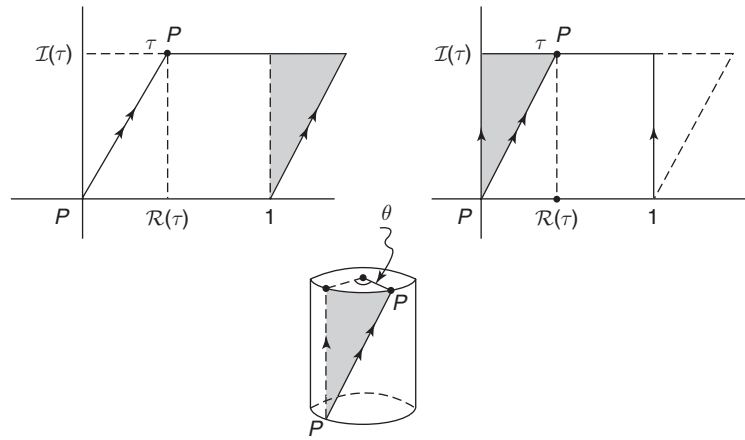

Fig. 26.18 **The canonical presentation of a torus with parameter $\tau \in \mathbb{H}$. The shaded region in the top left figure can be moved, as shown in the top right figure. Using this rectangular fundamental domain, we form a cylinder and recognize the twist angle θ associated with the final identification.**

Let us now explain the twisting of tori with $\Re(\tau) \neq 0$. Consider Figure 26.18. Let P denote the points $z = 0$ and $z = \tau$, which are equivalent after identification. Making use of the identification $z \sim z + 1$, we can move the shaded region, shown in the top left, to obtain the result shown in the top right. The fundamental domain has become a rectangle, even though the torus is not rectangular! While the vertical edges are identified, the identification of horizontal lines is shifted: the point P is not identified with a point that is directly above it, as would be the case for a rectangular torus. We can now roll up the rectangular domain to form a cylinder, whose ends remain to be joined together. This cylinder is shown in the bottom part of the figure. The identification of points on the two boundary components of the cylinder must be shifted, as P on the bottom end must be matched with P on the top end. We can use an angular variable running from zero to 2π to parameterize uniformly both boundary components of the cylinder. We take P on the lower boundary component to have zero angle. The upper boundary component is parameterized so that a given point takes the angular value assigned to the point on the lower boundary that is directly below it prior to identifications. Since the identifications are shifted by $\Re(\tau)$ along the boundary, and the boundaries have unit length, the angle θ associated with P on the top boundary is

$$\theta = 2\pi\, \Re(\tau)\,, \tag{26.27}$$

as shown in Figure 26.18. We call θ the twist angle.

We can now investigate the effects of letting the twist angle increase by 2π. This can be done by letting $\tau \to \tau + 1$, as can be seen in equation (26.27). If a torus with parameter τ and a torus with parameter $\tau + 1$ (Figure 26.19) are really the same, then increasing the twist angle by 2π changes nothing. We now use the arbitrariness in the choice of fundamental domain to demonstrate that two such tori are in fact the same. With the identification

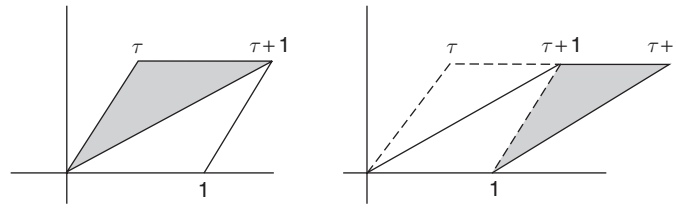

Fig. 26.19 **A torus with parameter τ (left) and a torus with parameter $\tau + 1$ (right). By moving the shaded region on the left using $z \sim z + 1$, we see that the two tori are really the same.**

$z \sim z + 1$, we can shift the shaded region on the left side of the figure to the position indicated on the right side of the figure. In doing so each point in the region has been moved horizontally by one unit. This operation does not change the torus; it is simply a different choice of fundamental domain. It follows that the torus on the left, with parameter τ, is the same torus as the torus on the right, with parameter $\tau + 1$:

$$\tau \sim \tau + 1\,. \tag{26.28}$$

Back in (26.27), this means that $\theta \sim \theta + 2\pi$. Thus, for example, twist angles beyond the interval $-\pi \leq \theta < \pi$ do not yield new tori. Now we can understand this concretely. Do *not* think of twisting as an application of physical torsion to wind up the cylinder before gluing the boundaries. Twisting is just a prescription for identifying points on the two boundary components of the cylinder: we arbitrarily pick a point P on the bottom boundary and then fix a point on the top boundary to identify it with. The identification is then extended uniformly over the boundaries. It is clear from this prescription that the twist identification parameter simply describes the position of a point on the top boundary, and thus the twist parameter lives on a circle.

Concerning the moduli space of tori, the identification (26.28) has an important implication. While τ certainly lives in $\mathbb{H}$, the space of inequivalent tori is much smaller. The identification shows that any infinite vertical strip of unit width contains all inequivalent tori. It is customary to choose the strip $\mathcal{S}_0$:

$$\mathcal{S}_0 \equiv \left\{ -\tfrac{1}{2} < \Re(\tau) \leq \tfrac{1}{2}\,, \quad \Im(\tau) > 0 \right\}. \tag{26.29}$$

The set of inequivalent tori is contained in $\mathcal{S}_0$. Note that the right boundary $\Re(\tau) = 1/2$ of the strip is included in the set. The left boundary $\Re(\tau) = -1/2$ is not included, because it is identified with the right boundary under $\tau \to \tau + 1$.

Our analysis of rectangular tori indicates that perhaps $\tau \sim -1/\tau$. Let us now prove that this is true. Consider a torus with parameter τ, as shown in the top left of Figure 26.20. Now define the conformally related $\tilde{z} = z/\tau$. This map is a rigid rotation plus a uniform scaling, and the original parallelogram turns into the parallelogram shown at the top right part of the figure. In a final step we do a rigid translation $z' = \tilde{z} - 1/\tau$. The final parallelogram in the z'-plane is in canonical form, and the parameter of the associated torus is $(-1/\tau)$. Thus $\tau \sim -1/\tau$.

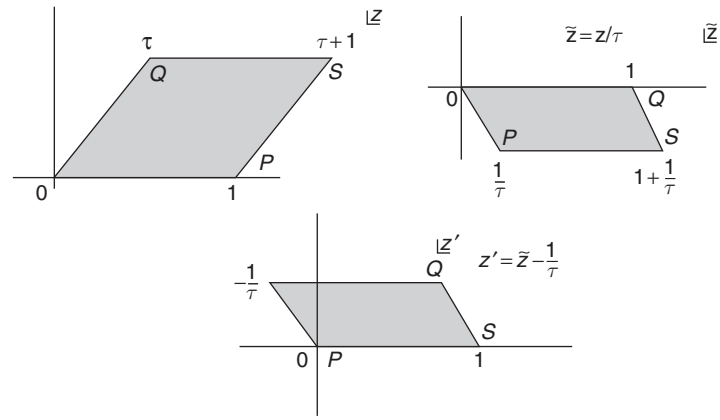

Fig. 26.20 **A sequence of conformal maps that shows that a torus with parameter τ is conformally equivalent to a torus with parameter $-1/\tau$.**

At this stage, we have found two equivalences for tori described by $\tau \in \mathbb{H}$:

$$\tau \sim \tau + 1\,, \quad \tau \sim -1/\tau\,. \tag{26.30}$$

It turns out that there are no additional independent identifications. It is therefore of great interest to find a fundamental domain for the identifications (26.30), for this will give us the set of all inequivalent tori. A fundamental domain for the first identification is the strip $\mathcal{S}_0$. Since the second relation identifies points in the region $|\tau| < 1$ with points in the region $|\tau| > 1$, we can attempt to use the subset of $\mathcal{S}_0$ which lies beyond the unit circle together with some points which lie on the unit circle. This is actually the answer, but it is by no means obvious. The same logic could have been used to argue that we can use the subset of $\mathcal{S}_0$ which lies inside the unit circle, together with some points which lie on the unit circle, but this would have been wrong. The claim is that a fundamental domain is the region $\mathcal{F}_0$, defined by

$$\mathcal{F}_0 \equiv \Big\{ -\tfrac{1}{2} < \Re(\tau) \le \tfrac{1}{2}\,, \Im(\tau) > 0,\ \ |\tau| \ge 1\,,\ \text{with the further restriction that}\ \ \Re(\tau) \ge 0 \text{ if } |\tau| = 1 \Big\}. \tag{26.31}$$

The fundamental domain $\mathcal{F}_0$ is shown as the shaded region in Figure 26.21. As stated in the above conditions, certain boundaries (shown by dashed lines in the figure) are not included in $\mathcal{F}_0$. These boundaries are composed of the points on the vertical line $\Re(\tau) = -1/2$ and by the points on the unit circle that lie to the left of $\tau = i$, which are removed by the further

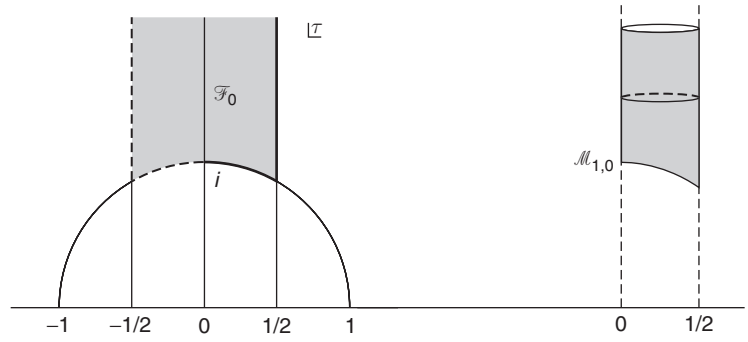

Fig. 26.21 **The fundamental domain $\mathcal{F}_0$ for the identifications $\tau \sim \tau + 1$ and $\tau \sim -1/\tau$ of $\mathbb{H}$. The closure $\bar{\mathcal{F}}_0$ of the region $\mathcal{F}_0$, with suitable identifications is the moduli space $\mathcal{M}_{1,0}$ of tori shown to the right.**

restriction in (26.31). We will give below part of the proof that $\mathcal{F}_0$ is a fundamental domain for the identifications (26.30), leaving the rest of the proof for Problem 26.6.

The moduli space of tori, denoted as $\mathcal{M}_{1,0}$, is the τ upper half-plane $\mathbb{H}$ subject to the identifications (26.30). We learned in Section 2.7 how to build a space which arises from identifications: we take the fundamental domain together with its boundary and implement the identifications on the boundary. In the present situation, the fundamental domain $\mathcal{F}_0$ together with its boundary defines the closure $\bar{\mathcal{F}}_0$:

$$\bar{\mathcal{F}}_0 = \left\{ -\tfrac{1}{2} \le \Re(\tau) \le \tfrac{1}{2}\,,\, \Im(\tau) > 0,\ \ |\tau| \ge 1 \right\}. \tag{26.32}$$

On $\bar{\mathcal{F}}_0$ we impose two identifications: $\Re(\tau) = -1/2$ and $\Re(\tau) = 1/2$ are identified by $\tau \to \tau + 1$, and the points with $|\tau| = 1$ are identified among themselves by $\tau \to -1/\tau$. These are, in fact, all the identifications in $\bar{\mathcal{F}}_0$ (see Problem 26.6). The resulting moduli space $\mathcal{M}_{1,0}$ of tori can be visualized by cutting out the region $\mathcal{F}_0$, folding it along the imaginary axis, and gluing the boundaries, as shown in the right side of Figure 26.21.

Quick calculation 26.1 Consider the points on the unit circle $|\tau| = 1$ that lie on $\mathbb{H}$ in between the vertical lines $\Re(\tau) = \pm 1/2$. Show that the points to the right of $\tau = i$ are identified with the points to the left of $\tau = i$ via $\tau \to -1/\tau$.

Let us consider the physical implications of this result. Since all inequivalent tori are contained in $\mathcal{F}_0$, one-loop closed string amplitudes must only include the contributions from tori with $\tau \in \mathcal{F}_0$. Since the tori in $\mathcal{F}_0$ are long tori, ultraviolet divergences are not an issue. If amplitudes used the full strip $\mathcal{S}_0$, the situation would have been quite problematic. This is not because we would have had some small tori to deal with, since, after all, they are conformally equivalent to (safe) long tori. The problem would have arisen because the complement of $\mathcal{F}_0$ in $\mathcal{S}_0$ contains an *infinite* number of copies of all the tori in $\mathcal{F}_0$! We can easily verify, for example, that it contains infinitely many copies of the torus $\tau = i$. Consider the set τ_n of values

$$\tau_n = i + n\,, \quad n \ge 1\,, \tag{26.33}$$

all of which, by virtue of $\tau \sim \tau + 1$, are just copies of the $\tau = i$ torus. Using the identification $\tau \sim -1/\tau$, we conclude that

$$-\frac{1}{\tau_n} = -\frac{n}{n^2+1} + \frac{i}{n^2+1} \tag{26.34}$$

are also copies of the $\tau = i$ torus. But for $n \geq 1$, these copies lie in $\mathcal{S}_0 - \mathcal{F}_0$. This is clear because the real part of $-1/\tau_n$ is in $[-1/2, 0]$ and $|1/\tau_n| = 1/\sqrt{n^2+1} < 1$. If we had to include all tori in $\mathcal{S}_0$, all one-loop amplitudes would be infinite. Instead, one-loop amplitudes use $\mathcal{F}_0$ because this space contains every inequivalent torus exactly once.

To establish that $\mathcal{F}_0$ is a fundamental domain for the identifications (26.30) requires some real work. Let us do some of it here to give you an idea about the necessary tools. To begin with, it is helpful to introduce some additional structure. Note that the identifications (26.30) are actually real linear fractional transformations, just like the ones we studied in Section 25.6. Let T and S denote the generators of linear fractional transformations that act as

$$T\,\tau = \tau + 1\,, \quad S\,\tau = -1/\tau\,. \tag{26.35}$$

Consider now transformations g of the form

$$g\tau = \frac{a\tau + b}{c\tau + d}\,, \quad a, b, c, d \in \mathbb{Z}\,, \quad ad - bc = 1\,. \tag{26.36}$$

Note that this is a linear fractional transformation with *integer* parameters. We can use matrix notation to describe these transformations. Associated with the transformation g we introduce the 2-by-2 matrix $[g]$ defined by

$$[g] = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\,, \quad \det[g] = 1\,. \tag{26.37}$$

The matrix $[g]$ has an ambiguity. We can change the sign of all entries without changing the transformation. We must therefore consider a matrix and minus the matrix to be equivalent. Given two transformations g_1 and g_2 of the form indicated in (26.36), the composition $g_1 g_2$ is also a transformation of the same form, and

$$[g_1 g_2] = [g_1]\,[g_2]\,. \tag{26.38}$$

Quick calculation 26.2 Prove equation (26.38).

The matrices associated with the transformations T and S are conventionally chosen to be

$$[T] = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\,, \quad [S] = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\,. \tag{26.39}$$

The composition of any number of S and T transformations, in any order, gives rise to a transformation of the type (26.36). A useful property of such transformations is that

$$\Im(g\tau) = \frac{\Im(\tau)}{|c\tau + d|^2}\,, \tag{26.40}$$

as we verified in a similar calculation earlier (see (25.66)).

The set of linear fractional transformations with integer coefficients (a, b, c, d) that satisfy $ad - bc = 1$ defines a group G of transformations under function composition. The group is called the *modular group*, and it is the same group as the group of 2-by-2 matrices with integer entries and unit determinant – with the proviso that a matrix and minus the matrix are considered to be the same group element. Here group multiplication is matrix multiplication. The modular group is called $PSL(2, \mathbb{Z})$. Here L stands for linear transformations or matrices. The two is for 2-by-2, the size of the matrices. S is for special, since the matrices have unit determinant. $\mathbb{Z}$ indicates that the entries are integers. Finally, P is for projective, or the fact that the sign of all entries can be reversed.

Let G' denote the set of projective transformations that can be generated by composition of any number of S and T transformations and their inverses. G' is clearly a subgroup of the modular group G. We will now prove that, for any $\tau \in \mathbb{H}$, there is a $g \in G'$ such that $g\tau \in \mathcal{F}_0$. This states that $\mathcal{F}_0$ contains a copy of each torus – a necessary condition for $\mathcal{F}_0$ to be a fundamental domain. The proof is not a trivial matter, the problem being that it is not all that easy to see how the tori in $\mathcal{S}_0 - \mathcal{F}_0$ can be sent into $\mathcal{F}_0$. The strategy requires making sure that we can send any point to a new point with a sufficiently large imaginary part, so that it can then be brought into $\mathcal{F}_0$ by some number of T transformations.

First we show that, for each τ, there is a $g \in G'$ such that $\Im(g\tau)$ is largest. It follows from (26.40) that we need only search for a g such that $|c\tau + d|$ is smallest. Since c and d are integers, as we vary them, $(c\tau + d)$ forms the lattice of points generated by τ and 1 in the complex plane. For any number $\alpha > 0$, there are at most a finite number of points such that $|c\tau + d| < \alpha$. Thus we should be able to find a minimum for $|c\tau + d|$ among all g, which gives a g for which $\Im(g\tau)$ is largest. The transformation g is not unique.

Now act with T^n on $g\tau$, with $n \in \mathbb{Z}$ chosen so that $T^n g\tau \in \mathcal{S}_0$. We claim that, unless $T^n g\tau \in \bar{\mathcal{F}}_0$, we have a contradiction. Let

$$\tau' = T^n \, g \, \tau \, . \tag{26.41}$$

The imaginary part of τ' is equal to the imaginary part of $g\tau$, and, by our proof above, it should not be possible to find a $g' \in G'$ such that $g'\tau'$ has an imaginary part larger than that of τ'. If τ' is not in $\bar{\mathcal{F}}_0$ then $|\tau'| < 1$. But then, using (26.40), we find that

$$\Im(S\tau') = \frac{\Im(\tau')}{|\tau'|^2} > \Im(\tau') \, , \tag{26.42}$$

contradicting the fact that no transformation in G' can increase the imaginary part of τ'. This contradiction proves that any point $\tau \in \mathbb{H}$ can be brought into $\bar{\mathcal{F}}_0$ by transformations generated by T and S (and their inverses). Once it is in $\bar{\mathcal{F}}_0$, then, if it is in $\mathcal{F}_0$ we are done. If it is not in $\mathcal{F}_0$, then it must lie in the boundary of $\bar{\mathcal{F}}_0$. Applying S or T once will then send it to $\mathcal{F}_0$. This completes the proof that any point $\tau \in \mathbb{H}$ can be brought into $\mathcal{F}_0$ by transformations generated by T and S.

Of course, more remains to be shown in order to prove that $\mathcal{F}_0$ is a fundamental domain. It must be established that no two points in $\mathcal{F}_0$ are connected via a transformation in G'. In fact, it is possible to show a stronger result: no two points are connected via a

transformation in the modular group G. Once this is proven, one can finally show that G' coincides with G. This means that S and T generate the full modular group (see Problem 26.6). It also follows that $\mathcal{F}_0$ is a fundamental domain for the modular group.

Problems

Problem 26.1 Open–closed string transition.

Examine the process by which an open string with light-cone momentum p^+ closes up to form a closed string.

(a) Draw the light-cone diagram as an infinite cylinder with a semi-infinite slit parallel to its axis. Show the incoming open string, the outgoing closed string, and the special closed string created at the interaction point.
(b) To draw this diagram as a region of the w plane, cut open the cylinder by following the direction of the slit. The result is the infinite strip $0 < \Im(w) \leq 2\pi\alpha' p^+$. The open strings exist for $\Re(w) < 0$, and the closed strings exist for $\Re(w) > 0$. The lines $\Im(w) = 0$ and $\Im(w) = 2\pi\alpha' p^+$ are glued when $\Re(w) > 0$. What is the boundary of the string diagram?
(c) Construct a map from this light-cone diagram into $z \in \bar{\mathbb{H}}$. Require that the incoming open string in the infinite past is mapped to $z = 0$ and that the outgoing closed string in the infinite future is mapped to $z = i$.
Hints: use $\xi = \exp(w/(2\alpha' p^+))$ to map the strip to $\bar{\mathbb{H}}$. Then let $\eta = \xi^2$. Why does this map remove the need for identifications? Show that the surface has gone to the exterior of a finite slit on the full η complex plane. To map this to $z \in \bar{\mathbb{H}}$ you will need a map that involves square roots.
(d) Plot in the z-plane the closed string emerging at the interaction time. Sketch also incoming open strings and outgoing closed strings.

Problem 26.2 Mapping of annuli into canonical form.

Find the error in the following argument which suggests that it is not possible to map conformally a topological annulus into a canonical annulus.

> The Riemann mapping theorem implies that any topological disk can be mapped to the unit disk. A topological annulus can therefore be mapped into the inside of a unit disk $|z| \leq 1$, with the outer boundary mapping into $|z| = 1$ and the inner boundary mapping into some closed curve inside the disk. To map this into a canonical annulus we must make the inner curve round and centered at the origin, while preserving the boundary $|z| = 1$. This requires a map of the disk to itself. The mappings of a disk to itself have three real parameters, just as the mappings of $\bar{\mathbb{H}}$ to itself. It is not possible to round a general curve using just three parameters, so it is impossible to map the annulus into canonical form.

Problem 26.3 Moduli space of T^2 compactifications with B field.

In Problem 17.5, the compactification parameters of the square torus T^2 are R and the Kalb–Ramond field b. Define the complex variable

$$\rho \equiv b\,\frac{R^2}{\alpha'} + i\,\frac{R^2}{\alpha'} \;=\; \frac{f_B}{2\pi} + i\,\frac{R^2}{\alpha'}\,.$$

Note that $\rho \in \mathbb{H}$. Prove that the transformations (1) and (2) in Problem 17.5 take the form

$$\rho \;\to\; \rho + 1\,, \quad \rho \;\to -1/\rho\,.$$

Describe the moduli space of compactifications on square tori T^2.

Problem 26.4 Shapiro–Virasoro amplitude.

The Shapiro–Virasoro amplitude is the scattering amplitude of four closed string tachyons. Recall that for a closed string tachyon we have

$$p^2 = -M^2 = 4/\alpha'\,. \tag{1}$$

For this amplitude we need the moduli space $\mathcal{M}_{0,4}$ of four-punctured spheres. When three of the punctures are placed at 0, 1, and ∞, the moduli space is parameterized by the position of the last puncture, a complex number $\lambda \in \hat{\mathbb{C}}$. The self-mappings of the z-sphere $z \in \hat{\mathbb{C}}$ are the linear fractional transformations

$$z \to \frac{az+b}{cz+d}\,, \quad ad - bc = 1\,. \tag{2}$$

(a) Suppose we want to integrate some quantity over the sphere. We then write an integral $\int d^2z \ldots$. Here $d^2z \equiv dxdy$, with $z = x + iy$. How does d^2z transform under a linear fractional transformation? (Recall that you must use Jacobians!)

(b) Mimicking the open string Veneziano amplitude, consider an expression for the Shapiro–Virasoro amplitude $A_{\rm SV}$ of the form

$$A_{\rm SV} = g^2 \int d^2z |z_1 - z_3|^{\alpha} \cdots |z - z_1|^{\beta p_2 \cdot p_1} \cdots \tag{3}$$

where the dots represent additional factors, which you must write, and α and β are constants you must determine by the condition that the integrand is invariant under linear fractional transformations.

(c) Simplify (3) now for the case $z_1 = 0$, $z_3 = 1$, $z_4 = \infty$, and $z = \lambda$. Show that you get an expression of the form

$$A_{\rm SV} = g^2 \int d^2\lambda\, |\lambda|^{\beta p_2 \cdot p_1} |1 - \lambda|^{\beta p_2 \cdot p_3}\,. \tag{4}$$

(d) To do the integral, first prove the following identity:

$$|z|^{-a} = \frac{1}{\Gamma(a/2)} \int_0^\infty dt\, t^{\frac{a}{2}-1} \exp(-t|z|^2)\,. \tag{5}$$

Now use it twice in (4), once for each factor in the integrand. Call the needed parameters t and s. The λ integral now becomes Gaussian, and this is made clear by writing $\lambda = \lambda_1 + i\lambda_2$ and $d^2\lambda = d\lambda_1 d\lambda_2$. Do the Gaussian integral. To do now the t and s integrals let $t = xu$ and $s = (1-x)u$ with $0 < x < 1$ and $0 < u < \infty$. Your final answer should be the ratio of three gamma functions over three other gamma functions.

(e) Consider the invariants

$$s = -(p_1 + p_2)^2\,, \quad t = -(p_2 + p_3)^2\,, \quad u = -(p_1 + p_3)^2\,.$$

Show that your answer has the expected pole structure in each of the three possible channels.

Problem 26.5 Bringing τ into the fundamental domain.

Let $\tau = (9 + i)/10$. Find the transformation in the modular group that brings it into a point τ_0 of the fundamental domain $\mathcal{F}_0$. Show that the transformation is unique by proving that there is no transformation in the modular group (except the identity) that leaves τ_0 fixed. Partial answer: $\tau_0 = 5i$. Repeat the above computations for $\tau = 9/10 + i/100$. Partial answer: $\tau_0 = i + 1/10$. [Hint: use the procedure discussed in the proof that any τ can be brought into the fundamental domain.]

Problem 26.6 S and T generate the modular group.

(a) Let $z \in \bar{\mathcal{F}}_0$ and assume that $gz \in \bar{\mathcal{F}}_0$ is a different point, where g is an element of the modular group (as in (26.37)). Since we can replace (z, g) by (gz, g^{-1}), we can assume that $\Im(gz) \geq \Im(z)$. This requires $|cz + d| \leq 1$. Explain why $|c| \geq 2$ is impossible. Discuss the finitely many possible values of c and d and prove that either $\Re(z) = \pm 1/2$ and $gz = z \mp 1$ or $|z| = 1$ and $gz = -1/z$.

(b) Prove that the group G' generated by S and T (and its inverses) coincides with the modular group G. Since G' is a subgroup of G, it suffices to show that each element of G is an element of G'. Construct the proof using the following two facts: (i) any point outside $\mathcal{F}_0$ can be brought into $\mathcal{F}_0$ by a transformation in G', and (ii) for any point z_0 on $\mathcal{F}_0$ and any $g \in G$ different from the identity, the point gz_0 is outside $\mathcal{F}_0$ (this is a consequence of (a)).

References

Most works in string theory assume background knowledge in quantum field theory and general relativity. For this reason, the present bibliography consists mostly of review papers. Even reviews are not an easy read, but they are more pedagogical than research papers, and they give many references, some of which the reader may want to consult at some point. We have also included in this bibliography some readable articles that supplement the material that has been covered in this book.

Readers of this book who are looking for more advanced material in string theory are well advised to begin their search by consulting the textbook by Becker, Becker, and Schwarz (2007), the two-volume text by Polchinksi (1998), and the two-volume text by Green, Schwarz, and Witten (1987). A helpful reference text is that of Kiritsis (2007). A detailed discussion of matters related to D-branes can be found in the book by Johnson (2003). While all these books require a background in quantum field theory and general relativity, selected parts may be understood without such a background.

For a view of string theory at its earliest stage of development, see the book by Frampton (1974). For string field theory, see the book by Siegel (1988). A book with a broad scope is that by Kaku (2000). Useful lecture notes include those of Lüst and Theisen (1989), Ooguri and Zin (1997), and 't Hooft (2004). A number of popular-level books are good reads: Gell-Mann (2002) for string theory within physics, Greene (2000) for an introduction to string theory, Randall (2005) for extra dimensions, and Susskind (2006) for string theory and the landscape.

The main reference for Part I of this book is the article of Goddard *et al.* (1973), which developed the light-cone quantization of the bosonic open string. After studying Part I, the reader may enjoy having a close look at this paper. Related subjects were nicely reviewed early on by Scherk (1975). The discussion in Section 7.4 is based on unpublished work of Jeffrey Goldstone. Background material for Part I can be found in textbooks on special relativity, electrodynamics, classical mechanics, and quantum mechanics. Although not needed, some students may enjoy Hartle's introduction to general relativity (2003) and Mandl and Shaw's introduction to quantum field theory (1984).

We now refer to articles and reviews that touch on specific topics. Problem 2.9 is based on a calculation by Seiberg (1997). For experiments testing the existence of large extra dimensions, see Chiaverini *et al.* (2003), Hoyle (2003), Long and Price (2003), and Kapner *et al.* (2007). For lectures on extra dimensions see Sundrum (2005). Problem (4.6) was based on a suggestion by Alan Guth. Problem 6.10, on a circular string in de Sitter space, is based on a homework problem by Frey (2005). Cosmic strings are discussed in detail in Vilenkin and Shellard (2000). For a status report see Polchinski (2006), and for cosmological bounds see Jeong and Smoot (2004). The calculation of the central term in the

Virasoro algebra follows a method by 't Hooft (2004). On the subject of tachyons and D-brane decays (the Sen conjectures), two references are Gaiotto and Rastelli (2003) and Sen and Zwiebach (2000). See also the reviews by Ohmori (2001), Sen (1999), and Taylor and Zwiebach (2003). The analytic description of the tachyon vacuum was given by Schnabl (2006). To supplement the discussion of superstrings in Chapter 14, the reader may consult the lectures by Schwarz (2000) and many other long and short reviews. A selected sample is: Douglas (1996), Polchinski (1996), Sagnotti and Sevrin (2002), Schwarz (2003), Schwarz and Seiberg (1999), Sen (2001), Taylor (2001), and Witten (2002).

Many of the subjects discussed in Part II are covered (more technically) in the books cited above. For D-branes, the reader may also consult the lectures by Bachas (1998), Polchinski, Chaudhuri, and Johnson (1996), Taylor (1998), the review by Giveon and Kutasov (1999), and the article by Hanany and Witten (1997). For some perspective on QCD within the Standard Model, see Gross (1999).

Electromagnetic duality is discussed in Deser (1982) and the review by Harvey (1996). For reviews on T-duality of closed strings, see Giveon, Porrati, and Rabinovici (1994) and Lerche, Schellekens, and Warner (1989). Open string T-duality and electromagnetic fields on D-branes are interrelated subjects. Some reviews are Ambjorn *et al.* (2003) and Taylor (1998). For additional details on strings in electric fields, see Seiberg, Susskind, and Toumbas (2002). For strings in magnetic fields, see Seiberg and Witten (1999) and the readable discussion of Bigatti and Susskind (2000a). A clear and detailed discussion of nonlinear electrodynamics is given by Bialynicki-Birula (1983). For references and results on Born–Infeld theory, see Gibbons (2003). Problems 20.6 and 20.7 are based on a result by Callan and Maldacena (1998).

The discussion of intersecting brane models and string phenomenology is based on the articles by Ibanez, Marchesano, and Rabadan (2001) and Cremades, Ibanez, and Marchesano (2002a, 2003). For reviews, see Uranga (2003) and Cremades, Ibanez, and Marchesano (2002b). Earlier works on the subject include Cvetic, Shiu, and Uranga (2001) and Blumenhagen *et al.* (2000). For other aspects of string theory model building, see the reviews by Aldazabal *et al.* (2000), Angelantonj and Sagnotti (2002), Dienes (1997), Ovrut (2002), and Quevedo (1996, 2002). Our discussion of moduli stabilization and the landscape borrows from Denef, Douglas, and Kachru (2007). Key ideas on this subject appeared in Bousso and Polchinski (2000). For a recent discussion of the landscape and particle physics see Lüst (2007).

An insighful historical introduction to thermodynamics and statistical mechanics is found in the book by Atkins (1994). Since the subject of string thermodynamics is very broad, we cite Abel *et al.* (1999), a relatively recent paper with many references. Also recommended are the lectures by Deo, Jain, and Tan (1990). For a readable introduction to the physics of black holes see, for example, the book by Begelman and Rees (1998) and the book by Thorne (1994). See also the article by Bekenstein (2001). Das and Mathur (2001) review many aspects of black hole physics in string theory. Our discussion of the entropy of Schwarzschild black holes is based on Horowitz and Polchinski (1997). The five-dimensional black hole discussed in the text is that in Strominger and Vafa (1996). For a novel approach to black hole entropy see Mathur (2005). For a recent review on the subject see Sen (2007).

For Regge trajectories see Tang and Norbury (2000) and Selem and Wilczek (2006). Problem 23.1 is based on a similar problem in Frey (2005). The Poincaré-invariant modification of string theory with critical dimension four is in Polchinski and Strominger (1991). For a review of results on the quark-antiquark potential see Kuti (2005). For the AdS/CFT correspondence, a readable, brief review is Horowitz and Polchinski (2006). Other useful reviews are: Aharony *et al.* (2000), D'Hoker and Freedman (2002), Klebanov (2000), Maldacena (2003), Mateos (2007), and Petersen (1999). The AdS/CFT correspondence originated from the Maldacena conjecture (1998). For other reviews on related topics, see Banks (1999), Bigatti and Susskind (1997, 2000b), Taylor (2001), and Witten (1998). On the quark-gluon plasma, see the popular-level accounts by Riordan and Zajc (2006) and by Blau (2005). Recent technical references are Son and Starinets (2007), Liu, Rajagopal, and Wiedemann (2006), and Herzog *et al.* (2006).

Covariant quantization and string interactions are discussed in the string theory textbooks cited before. Readers who wish to go beyond our basic discussion of the Virasoro operators may look at the lecture notes by Ginsparg (1991), which give an efficient introduction to conformal field theory. A readable book on Riemann surfaces is that by Springer (1981). For a construction of all Riemann surfaces without short possible strings, see Zwiebach (1993). Our discussion of the modular group follows the book by Serre (1973).

References

Abel, S. A., Barbon, J. L., Kogan, I. I., and Rabinovici, E. (1999). String thermodynamics in D-brane backgrounds, *J. High Energy Phys.* **9904**, 015 [arXiv:hep-th/9902058].

Aharony, O., Gubser, S. S., Maldacena, J. M., Ooguri, H., and Oz, Y. (2000). Large N field theories, string theory and gravity, *Phys. Rep.* **323**, 183 [arXiv:hep-th/9905111].

Aldazabal, G., Ibanez, L. E., Quevedo, F., and Uranga, A. M. (2000). D-branes at singularities: a bottom-up approach to the string embedding of the standard model, *J. High Energy Phys.* **0008**, 002 [arXiv:hep-th/0005067].

Ambjorn, J., Makeenko, Y. M., Semenoff, G. W., and Szabo, R. J. (2003). String theory in electromagnetic fields, *J. High Energy Phys.* **0302**, 026 [arXiv:hep-th/0012092].

Angelantonj, C. and Sagnotti, A. (2002). Open strings, *Phys. Rep.* **371**, 1 [Erratum **376**, 339 (2003)] [arXiv:hep-th/0204089].

Atkins, P. W. (1994). *The Second Law*. New York: Scientific American Library.

Bachas, C. P. (1998). Lectures on D-branes, arXiv:hep-th/9806199.

Banks, T. (1999). TASI lectures on matrix theory, arXiv:hep-th/9911068.

Bartlett, J. (1919). *Familiar Quotations*. Boston: Little, Brown, and Company.

Becker, K., Becker, M., and Schwarz, J. H. (2007) *String Theory and M-Theory.* Cambridge: Cambridge University Press.

Begelman, B. and Rees, M. (1998). *Gravity's Fatal Attraction*. New York: Scientific American Library.

Bekenstein, J. D. (2001). The limits of information, *Stud. Hist. Philos. Mod. Phys.* **32**, 511 [arXiv:gr-qc/0009019].

Bialynicki-Birula, I. (1983). Non-linear electrodynamics: variations on a theme of Born and Infeld. In *Quantum Theory of Fields and Particles*, ed. B. Jancerwicz and J. Lukierski. Singapore: World Scientific.

Bigatti, D. and Susskind, L. (1997). Review of matrix theory, arXiv:hep-th/9712072.

(2000a). Magnetic fields, branes and noncommutative geometry, *Phys. Rev. D* **62**, 066004 [arXiv:hep-th/9908056].

(2000b). TASI lectures on the holographic principle, arXiv:hep-th/0002044.

Blau, S. K. (2005) A string-theory calculation of viscosity could have surprising applications. *Phys. Today* **58N5**:23.

Blumenhagen, R., Goerlich, L., Kors, B., and Lüst, D. (2000). Noncommutative compactifications of type I strings on tori with magnetic background flux, *J. High Energy Phys.* **0010**, 006 [arXiv:hep-th/0007024].

Bousso, R. and Polchinski, J. (2007). Quantization of four form fluxes and dynamical neutralization of the cosmological constant. *J. High Energy Phys.* **0006**:006 [arXiv: hep-th/0004134].

Callan, C. G. and Maldacena, J. M. (1998). Brane dynamics from the Born–Infeld action, *Nucl. Phys. B* **513**, 198 [arXiv:hep-th/9708147].

Chiaverini, J., Smullin, S. J., Geraci, A. A., Weld, D. M., and Kapitulnik, A. (2003). New experimental constraints on non-Newtonian forces below 100-mu-m, *Phys. Rev. Lett.* **90**, 151101 [arXiv:hep-ph/0209325].

Cremades, D., Ibanez, L. E., and Marchesano, F. (2002a). Intersecting brane models of particle physics and the Higgs mechanism, *J. High Energy Phys.* **0207**, 022 [arXiv:hep-th/0203160].

(2002b). More about the Standard Model at intersecting branes, arXiv:hep-ph/0212048.

(2003). Yukawa couplings in intersecting D-brane models, *J. High Energy Phys.* **0307**, 038 [arXiv: hep-th/0302105].

Cvetic, M., Shiu, G., and Uranga, A. M. (2001). Three-family supersymmetric standard like models from intersecting brane worlds, *Phys. Rev. Lett.* **87**, 201801 [arXiv:hep-th/0107143].

Das, S. R. and Mathur, S. D. (2001). The quantum physics of black holes: results from string theory, *Annu. Rev. Nucl. Sci.* **50**, 153 [arXiv:gr-qc/0105063].

Denef, F., Douglas, M. R., and Kachru, S. (2007). Physics of String Flux Compactifications, *Ann. Rev. Nucl. Part. Sci.* **57**, 119 [arXiv: hep-th/0701050].

Deo, N., Jain, S., and Tan, C. I. (1990). The ideal gas of strings, *Bombay Quant. Field Theory* **1990** 112–148 (scanned version: KEK library).

Deser, S. (1982). Off-shell electromagnetic duality invariance, *J. Phys. A: Math. Gen*. **15** 1053.

D'Hoker, E. and Freedman, D. Z. (2002). Supersymmetric gauge theories and the AdS/CFT correspondence, arXiv:hep-th/0201253.

Dienes, K. R. (1997). String theory and the path to unification: a review of recent developments, *Phys. Rep.* **287**, 447 [arXiv:hep-th/9602045].

Douglas, M. R. (1996). Superstring dualities, Dirichlet branes and the small scale structure of space, arXiv:hep-th/9610041.

Frampton, P. H. (1974). *Dual Resonance Models*. Frontiers in Physics Series. Benjamin.

Frey, A. (2005). http://theory.caltech.edu/∼frey/ph135/.

Gaiotto, D. and Rastelli, L. (2003). Experimental string field theory, *J. High Energy Phys.* **0308**, 048 [arXiv:hep-th/0211012].

Gell-Mann, M. (2002). *The Quark and the Jaguar*. New York: Henry Holt and Company, LLC.

Gibbons, G. W. (2003). Aspects of Born–Infeld theory and string/M-theory, *Rev. Mex. Fis.* **49S1**, 19 [arXiv:hep-th/0106059].

Ginsparg, P. (1991). Applied conformal field theory, arXiv:hep-th/9108028.

Giveon, A. and Kutasov, D. (1999). Brane dynamics and gauge theory, *Rev. Mod. Phys.* **71**, 983 [arXiv:hep-th/9802067].

Giveon, A., Porrati, M., and Rabinovici, E. (1994). Target space duality in string theory, *Phys. Rep.* **244**, 77 [arXiv:hep-th/9401139].

Goddard, P., Goldstone, J., Rebbi, C., and Thorn, C. B. (1973). Quantum dynamics of a massless relativistic string, *Nucl. Phys. B* **56**, 109.

Green, M. B., Schwarz, J. H., and Witten, E. (1987). *Superstring Theory.* Vol. 1: *Introduction*, Vol. 2: *Loop Amplitudes, Anomalies and Phenomenology*. Cambridge: Cambridge University Press.

Greene, B. (2000). *The Elegant Universe*. New York: Vintage Books.

Gross, D. J. (1999). Twenty five years of asymptotic freedom, *Nucl. Phys. Proc. Suppl.* **74**, 426 [arXiv:hep-th/9809060].

Hanany, A. and Witten, E. (1997). Type IIB superstrings, BPS monopoles, and three-dimensional gauge dynamics, *Nucl. Phys. B* **492**, 152 [arXiv:hep-th/9611230].

Hartle, J. B. (2003). *Gravity*. San Francisco, CA: Addison-Wesley.

Harvey, J. A. (1996). Magnetic monopoles, duality, and supersymmetry, arXiv:hep-th/9603086.

Herzog, C. P., Karch, A., Kovtun, P., Kozcaz, C., and Yaffe, L. G. (2006). Energy loss of a heavy quark moving through N = 4 supersymmetric Yang-Mills plasma, *JHEP* **0607**, 013 [arXiv:hep-th/0605158].

Horowitz, G. T. and Polchinski, J. (1997). A correspondence principle for black holes and strings, *Phys. Rev. D* **55**, 6189 [arXiv:hep-th/9612146].

(2006). Gauge/gravity duality, arXiv: gr-qc/0602037.

Hoyle, C. D. (2003). The weight of expectation, *Nature* **421**, 899.

Ibanez, L. E., Marchesano, F., and Rabadan, R. (2001). Getting just the Standard Model at intersecting branes, *J. High Energy Phys.* **0111** 002 [arXiv:hep-th/0105155].

Jeong, E. and Smoot, G. F. (2005). Search for cosmic strings in CMB anisotropies, *Astrophys. J.* **624**, 21 [astro-ph/0406432].

Johnson, C. V. (2003). *D-branes*. Cambridge: Cambridge University Press.

Kaku, M. (2000). *Strings, Conformal Fields and M-theory*. New York: Springer.

Kapner, D. J., Cook, T. S., Adelberger, E. G., Gundlach, J. H., Heckel, B. R., Hoyle, C. D., and Swanson, H. E. (2007). Tests of the gravitational inverse-square law below the dark-energy length scale, *Phys. Rev. Lett.* **98**, 021101 [arXiv:hep-ph/0611184].

Kiritsis, E. (2007). *String Theory in a Nutshell*. Princeton, New Jersey: Princeton University Press.

Klebanov, I. R. (2000). TASI lectures: introduction to the AdS/CFT correspondence, arXiv:hep-th/0009139.

Kuti, J. (2005). Lattice QCD and string theory, arXiv: hep-lat/0511023.

Lerche, W., Schellekens, A. N., and Warner, N. P. (1989). Lattices and strings, *Phys. Rep.* **177**, 1.

Liu, H., Rajagopal, K., and Wiedemann, U. A. (2006). Calculating the jet quenching parameter from AdS/CFT, *Phys. Rev. Lett.* **97**, 182301 [arXiv:hep-ph/0605178].

Long, J. C. and Price, J. C. (2003). Current short-range tests of the gravitational inverse square law, arXiv:hep-ph/0303057.

Lüst, D. (2007) String Landscape and the Standard Model of Particle Physics, arXiv: hep-th/0707.2305.

Lüst, D. and Theisen, S. (1989). *Lectures on String Theory*. Berlin: Springer Verlag.

Maldacena, J. M. (1998). The large N limit of superconformal field theories and supergravity, *Adv. Theor. Math. Phys.* **2**, 231 (*Int. J. Theor. Phys.* **38**, 1113 (1999)) [arXiv:hep-th/9711200].

(2003). TASI 2003 lectures on AdS/CFT, arXiv:hep-th/0309246.

Mandl, F. and Shaw, G. (1984). *Quantum Field Theory*. New York: Wiley.

Mateos, D. (2007). String Theory and Quantum Chromodynamics, arXiv: hep-th/0709.1523.

Mathur, S. D. (2005). The Fuzzball proposal for black holes: An Elementary review. *Fortsch. Phys.* **53**, 793 [arXiv: hep-th/0502050].

Ohmori, K. (2001). A review on tachyon condensation in open string field theories, arXiv:hep-th/0102085.

Ooguri, H. and Zin, Y. (1997). Lectures on perturbative string theories. In *Fields, Strings, and Duality*, *TASI 1996*, ed. C. Efthimiou and B. Greene, pp. 5–82. Singapore: World Scientific [arXiv:hep-th/9612254].

Ovrut, B. A. (2002). Lectures on heterotic M-theory, arXiv:hep-th/0201032.

Petersen, J. L. (1999). Introduction to the Maldacena conjecture on AdS/CFT, *Int. J. Mod. Phys. A* **14**, 3597 [arXiv:hep-th/9902131].

Polchinski, J. (1996). String duality: a colloquium, *Rev. Mod. Phys.* **68**, 1245 [arXiv: hep-th/9607050].

(1998). *String Theory.* Vol. 1: *An Introduction to the Bosonic String*, Vol. 2: *Superstring Theory and Beyond*. Cambridge: Cambridge University Press.

(2006) Cosmic String Loops and Gravitational Radiation, arXiv:0707.0888.

Polchinski, J., Chaudhuri, S., and Johnson, C. V. (1996). Notes on D-branes, arXiv: hep-th/9602052.

Polchinski, J. and Strominger, A. (1991). Effective string theory, *Phys. Rev. Lett.* **67**, 1681.

Quevedo, F. (1996). Lectures on superstring phenomenology, arXiv: hep-th/9603074.

(2002). Lectures on string/brane cosmology, *Class. Quant. Grav.* **19**, 5721 [arXiv: hep-th/0210292].

Riordan, M. and Zajc, W. (2006). The first few microseconds, *Sci. Am.* **294N5**, 24.

Randall, L. (2005). *Warped Passages*. New York: Harper Collins.

Sagnotti, A. and Sevrin, A. (2002). Strings, gravity and particle physics, arXiv: hep-ex/0209011.

Scherk, J. (1975). An introduction to the theory of dual models and strings, *Rev. Mod. Phys.* **47**, 123.

Schnabl, M. (2006). Analytic solution for tachyon condensation in open string field theory, *Adv. Theor. Math. Phys.* **10**, 433 [arXiv: hep-th/0511286].

Schwarz, J. H. (2000). Introduction to superstring theory, arXiv:hep-ex/0008017.

(2003). Update on string theory, arXiv:astro-ph/0304507.

Schwarz, J. H. and Seiberg, N. (1999). String theory, supersymmetry, unification, and all that, *Rev. Mod. Phys.* **71**, S112 [arXiv:hep-th/9803179].

Seiberg, N. (1997). Why is the matrix model correct?, *Phys. Rev. Lett.* **79**, 3577 [arXiv: hep-th/9710009].

Seiberg, N. and Witten, E. (1999). String theory and noncommutative geometry, *J. High Energy Phys.* **9909**, 032 [arXiv:hep-th/9908142].

Seiberg, N., Susskind, L., and Toumbas, N. (2002). Strings in background electric field, space/time noncommutativity and a new noncritical string theory, *J. High Energy Phys.* **0006**, 021 [arXiv:hep-th/0005040].

Selem, A. and Wilczek, F. (2006). Hadron systematics and emergent diquarks, arXiv: hep-ph/0602128.

Sen, A. (1999). Non-BPS states and branes in string theory, arXiv:hep-th/9904207.

(2001). Recent developments in superstring theory, *Nucl. Phys. Proc. Suppl.* **94**, 35 [arXiv:hep-lat/0011073].

(2005). Tachyon dynamics in open string theory, *Int. J. Mod. Phys.* **A20**:5513 [arXiv: hep-th/0410103].

(2007). Black Hole Entropy Function, Attractors and Precision Counting of Microstates, arXiv:0708.1270.

Sen, A. and Zwiebach, B. (2000). Tachyon condensation in string field theory, *J. High Energy Phys.* **0003**, 002 [arXiv:hep-th/9912249].

Serre, J. P. (1973). *A Course in Arithmetic*. Graduate Texts in Mathematics. Berlin: Springer-Verlag.

Siegel, W. (1988). Introduction to string field theory, arXiv:hep-th/0107094.

Son, D. T. and Starinets, A. O. (2007). Viscosity, Black Holes, and Quantum Field Theory, *Ann. Rev. Nucl. Part. Sci.* **57**, 95 [arXiv:0704.0240].

Springer, G. (1981). *Introduction to Riemann Surfaces*. New York: Chelsea Publishing Company.

Strominger, A. and Vafa, C. (1996). Microscopic origin of the Bekenstein–Hawking entropy, *Phys. Lett. B* **379**, 99 [arXiv:hep-th/9601029].

Sundrum, R. (2005). TASI 2004 lectures: To the fifth dimension and back, arXiv: hep-th/0508134.

Susskind, L. (2006). *The Cosmic Landscape*. New York: Little, Brown and Company.

Tang, A. and Norbury, J. W. (2000). Properties of Regge trajectories, *Phys. Rev.* **D62**, 016006 [arXiv: hep-ph/0004078].

Taylor, W. (1998). Lectures on D-branes, gauge theory and M(atrices), arXiv: hep-th/9801182.

(2001). M(atrix) theory: matrix quantum mechanics as a fundamental theory, *Rev. Mod. Phys.* **73**, 419 [arXiv:hep-th/0101126].

Taylor, W. and Zwiebach, B. (2003). D-branes, tachyons, and string field theory, arXiv:hep-th/0311017.

Thorne, K. S. (1994). *Black Holes and Time Warps*. New York: Norton & Company.

't Hooft, G. (2004). Introduction to string theory. http://www.phys.uu.nl/~thooft/lectures/stringnotes.pdf

Uranga, A. M. (2003). Chiral four-dimensional string compactifications with intersecting D-branes, *Class. Quant. Grav.* **20**, S373 [arXiv:hep-th/0301032].

Vilenkin, A. and Shellard, E. P. S. (2000). *Cosmic Strings and Other Topological Defects*. Cambridge: Cambridge University Press.

Witten, E. (1998). New perspectives in the quest for unification, arXiv:hep-ph/9812208.
(2002). Comments on string theory, arXiv:hep-th/0212247.

Zwiebach, B. (1993). Closed string field theory: an introduction, arXiv:hep-th/9305026.

Index

action
  charged point particle, 98
  fields and symmetries, 157
  first defined, 78
  free particle in curved space, 99
  free point particle, 91
  free scalar field, 196
  Nambu–Goto, 111–112, 123, 178
  open string and charged endpoints, 368
  particle and EM field, 99, 356
  Polyakov string, 582–587
  saddle points, 87
  string and Kalb–Ramond field, 357
  transverse string coordinates, 248
admissible states, 575
AdS/CFT correspondence
  conformal field theory, 537
  finite temperature, 554–559
  holography, 553–554
  large N limit, 533–535
  motivating, 537–541
  near horizon geometry, 539
  parameters, 541–543
anti-de Sitter
  black hole, 555
  conformal boundary, 552
  defined, 549
  isometries, 537
  metric, 552, 564, 565

background fields, 361–362, 415–416
black hole
  in anti-de Sitter space, 555
  basics, 513–514
  built with D-branes, 519–520
  entropy, 514–515, 520–521
  in IIB superstrings, 518–519
  temperature, 514
Bohm–Aharonov effect, 406
Born–Infeld electrodynamics
  and T-duality, 443–446
  capacitance, 446
  energy of a point charge, 441–443, 447
  general theory, 438–443
  Hamiltonian, 447
  Lagrangian density, 438
  Lagrangian for Dp-brane, 445
  string ending on D-brane, 448–450

Calabi–Yau space, 479
capacitance
  and moduli of annuli, 636–641
  Born–Infeld electrodynamics, 446
Cauchy–Riemann equations, 599
Chan–Paton labels, 345
charge
  conserved
    for Lagrangian densities, 158
    Lagrangian mechanics, 156
    Lorentz, 167
    momentum, 160, 164
  electric, 154–155
chiral fermions, 463–464, 468, 472
conformal boundary
  AdS, 552
  $\mathbb{H}_n$, 548
  defined, 547
  Minkowski space, 564
  $\mathbb{R}_2$, 548
conformal gauge, 586
conformal map, *see* Schwarz–Christoffel
constraints
  from the Polyakov action, 587
  parameterization, 136, 138, 182, 569
  quantum Virasoro, 572–577
  solving in light-cone gauge, 187–189
cosmic strings, 145–149
counting
  symmetric products, 277
  with generating functions, 317–320
cross section, 595
current
  conserved
    defined, 154
    for Lagrangian densities, 158
    Lorentz, 167
    momentum, 159–161
  electric, 154–155
  four-vector, 50

cusp
in closed strings, 144, 152
in open strings, 151

D-branes
charge, 370–373
classical motion of endpoints, 128
classical open string motion, 149–151
Dp-brane boundary conditions, 331–333
D0-branes, 589
dissolved, 429, 431
first discussed in detail, 115–116
instability, 270–274
Lagrangian with EM fields, 445
open strings between Dp- and Dq-branes, 345–350
open strings between parallel Dp-branes, 338–345
open strings on Dp-branes, 333–338
tension T_p of Dp-brane, 413, 414
tilted, 424, 425
with electric fields, 418–422
with magnetic fields, 423–429
de Sitter space, 566
deficit angle, 145
degrees of freedom
defined, 199
gravity field, 212
Maxwell field, 208
descendent, *see* Virasoro descendent
dilaton
from covariant closed string, 590
from quantum closed string, 293
related to the string coupling, 294–296
dimensional reduction, 47, 389–390
Dirichlet boundary condition
defined, 75
open strings, 114
reformulated, 83
divergence
in higher dimensions, 70
theorem, 56–57
double strike, 303, 386
dual field strength $\tilde{F}^{\mu\nu}$, 437, 447
duality
defined, 376
for a classical oscillator, 377–378
in electromagnetism, 376–377
T-duality, *see* T-duality
Dynkin label, 466, 467

Einstein
general relativity, 58
summation convention, 17
electric field
critical, 422
motion of open strings, 429–430
on D-branes, 418–422
open string boundary conditions, 419
electromagnetism
gauge potentials, 48
current vector, 50
duality invariance, 376–377
field strength defined, 48
gauge transformations, 46
Maxwell equations, 45
ensemble
canonical, 495–498
microcanonical, 495–498
entropy
black hole, 514–515, 520–521
defined, 495
of a string, 506, 515–516
extra dimensions
Lorentz invariance, 30
compact, 31
large, 67–69

Feynman diagram, 592–594, 630
field strength, *see* electromagnetism
fine-structure constant, 62, 294
flux compactification, 484
four-vector, 21
free boundary condition, 114
fundamental domain
defined, 32
modular group $\mathcal{F}_0$, 651–656
modular group, exercise, 658
modular group, identifications, 658

gamma function
analytic continuation, 70
defined, 54
recursion relation, 54
Veneziano amplitude, 625
gauge transformations
$U(1)$ structure, 404–406
gravitational field, 60, 210
Kalb–Ramond field, 214
Maxwell and Kalb–Ramond, 368
Maxwell field, 46, 206
geodesic
equation, 99
on a cone, 147, 153
graviton states
from covariant closed strings, 590
from gravity field, 212
from light-cone closed strings, 292
gravity field
degrees of freedom, 212
gauge transformations, 210
light-cone gauge, 211
quantum theory, 209–213
group
$U(1)$, 344, 405
$U(N)$, 344

Hagedorn temperature
  bosonic string theory, 505–507
  open superstring theory, 522
Hamilton's principle, 79
Hamiltonian
  Born–Infeld electrodynamics, 447
  charged point particle, 99
  covariant open string, 570, 578
  density for scalar field, 197
  density for strings, 127
  energy for strings, 133
  light-cone closed string, 286, 289, 388
  light-cone open string, 238, 262
  light-cone point particle, 222
  point particle energy, 92
Hardy–Ramanujan, 502
Hawking–Page transition, 559
helicity, 463
hierarchy problem, 61, 69
Higgs mechanism, 345, 463, 465, 493
holography, 549
holonomy W
  as angle variable, 410–412
  defined, 408
  Wilson line, *see* Wilson line
hyperbolic space
  $\mathbb{H}_2$, 544
  $\mathbb{H}_2$ isometric embedding in $\mathbb{R}^3$, 562
  $\mathbb{H}_n$ defined, 545
  $\mathbb{H}_n$ metric, 547
  $\mathbb{H}_n$ conformal boundary, 548

intersection number, 454, 457, 491
interval, 15

jumping rope
  angular momentum, 172
  relativistic, 150

Kalb–Ramond field
  T^2 compactification, 398–399
  analysis, 214–215
  coupling to a string, 357
  coupling to Maxwell field on D-brane, 368–370
  electric charge density, 359
  gauge transformations, 214
  motion of a string, 374–375
  of a string, 374
  particle states, 215
  states from light-cone closed string, 292
Kaluza–Klein, 390

Lagrangian
  Born–Infeld electrodynamics, 438
  density for nonrelativistic string, 81
  first defined, 78
  for Dp-brane with EM fields, 445
  nonrelativistic string, 81
  symmetry
    defined, 155
    more general, 159, 173–174
landscape, 10, 482, 489
light-cone
  components of tensors, 213
  coordinates, 22
  energy and momentum, 28
  gauge
    gravity field, 211
    Maxwell field, 207
    open and closed strings, 187
    point particle, 217
  Hamiltonian, *see* Hamiltonian
  metric, 24
  scalar field equation, 200
lightlike
  compactification, 42
  separated events, 16
linear fractional transformation
  constructed from three points, 611, 626
  defined, 610
  number of parameters, 611
Lorentz algebra, 231, 234
Lorentz force law, 46, 69
Lorentz generators
  quantum open string, 259–262
  quantum point particle, 229–233
Lorentz transformations
  boost, 19
  defined, 20
  infinitesimal, 165
Luscher coefficient, 532

M-theory, 325, 481
magnetic field
  flux quantization, 427, 430
  motion of open string, 431–432
  on D-branes, 423–429
  related to tilting angle, 425
mass-squared
  closed string, 288
  closed string with compactification, 387
  DD contribution, 336
  ND contribution, 350
  normal ordering contributions, 349
  NS sector, 314
  open string, 263
  open string (classical), 190
  open string between D-branes, 341
  R sector, 316
  twisted sector, 302
massive vector field
  formulation, 215
  from separated D-branes, 342–343

Maxwell field
coupling to open strings, 368, 415–418
gauge transformations, 206
holonomy, 408
light-cone gauge, 207
photon states, 209
quantum theory, 206–209
metric
conformal, 586
dynamical, 59
for open string in magnetic field, 432
induced on surface, 104
induced on world-sheet, 111
Minkowski, 17–18
transformation under reparameterization, 105
world-sheet $h_{\alpha\beta}$, 583
mode expansion
closed string coordinate
no compactification, 284
with compactification, 383
for X^- open string coordinate, 190
for DD open string coordinate, 335
for ND open string coordinate, 348
for NN coordinate derivatives, 186
for NN open string coordinate, 186
modular group $PSL(2,\mathbb{Z})$
defined, 655
fundamental domain $\mathcal{F}_0$, 656
generated by S and T, 658
moduli space
of S^1 compactification, 393
of T^2 compactification with B field, 656
of Riemann surfaces, *see* Riemann surface
moduli stabilization, 481–484

Nambu–Goto
free string equations of motion, 113
string action, 111–112, 170, 178
Neumann boundary condition, 76
Neveu–Schwarz fermions, 312–315
Newton
constant G, 60
constant in arbitrary dimension, 63
constant in terms of g, 294
gravitational law, 60
nonlinear electrodynamics
Born–Infeld theory, 438–443
energy functional, 436, 446
field $\bar{D}$, 436
field $\bar{H}$, 436
general framework, 433–437
non-planar open string diagrams, 642–643
normal ordering, 251
null
states, 577
vector, 21
vectors orthogonal to, 191

number operator
closed string, 287
open string, 264

operator
creation and annihilation
closed string, 285
open string, 244–246
quantum scalar field, 203
Heisenberg, 218–220
open string Heisenberg, 237
open string Schrödinger, 237
point particle
Heisenberg, 221
Schrödinger, 220
scalar field operator, 204
Schrödinger, 218–220
orbifold
$T^2/\mathbb{Z}_3$, 41
$\mathbb{C}/\mathbb{Z}_N$, 35
$\mathbb{R}^1/\mathbb{Z}_2$
closed strings, 296–298
defined, 35
twisted sector, 298–303
construction, 41
defined, 35
spacetime, 43
orientation
of an open string, 339
orientifold plane
and coincident Dp-branes, 352
and Dp-brane, 351
and separated Dp-branes, 352–354
closed string spectrum, 305

particle states
Kalb–Ramond states, 215
one-graviton states, 212
one-particle, 205–206
one-photon states, 209
role in quantum field theory, 201
partition function
defined, 496
relativistic particle, 507–509, 522
single string, 509–513
partitions of N
$P(N;b,f)$, 505, 522
$p(N)$, 499, 502
$p_b(N)$, 503–504
$q(N)$, 504, 521
generating functions, 522
photon states
from covariant open string, 579–580, 589
from light-cone open string, 266–267
from Maxwell field, 209
physical state
covariant quantization

defined, 577
gravitons, 590
photons, 580
tachyons, 579
light-cone quantization
open string, 266
point particle, 224
Planck
length, mass, time, 61
constant, 36
energy, 62
length in arbitrary dimension, 63
point particle
action if charged, 98
coupled to dynamical EM field, 99
covariant quantization, 587
free action, 91
free equation of motion, 95
free motion in curved space, 99
Hamiltonian, when charged, 98
Heisenberg operators, 221
Lagrangian when free, 92
light-cone Lorentz generators, 229–233
light-cone momentum generators, 226–229
quantum states, 223
reparameterization symmetry, 234
Schrödinger equation, 224
Schrödinger operators, 220
primary, *see* Virasoro primary
proper time, 27
as action, 91–92
puncture, 602
pure gauge
defined, 208
gravity field, 214
Maxwell field, 209
states, 577

quantum field theory
gravity field, 209–213
Kalb–Ramond field, 214–215
massive vector field, 215
Maxwell field, 206–209
scalar field, 200–206
quark-antiquark potential, 531–532
quark-gluon plasma
jet-quenching, 561
production, 559
viscosity, 560

Ramond fermions, 315–317
Ramond–Ramond
couplings to D-branes, 371
fields in closed superstrings, 324
Regge trajectory, 527
reparameterization
exercise for particle, 98
generated by Virasoro, 257–259, 277, 289–290
invariance for particle, 93–94, 234
invariance of area, 103–106
invariance of Kalb–Ramond coupling, 373
Riemann surface
defined, 599
moduli space
$\mathcal{M}_{0,3}$, 616
$\mathcal{M}_{0,4}$, 611–613, 646
$\mathcal{M}_{1,0}$, 652
$\mathcal{N}_3$, 616
$\mathcal{N}_4$, 616–617
of annuli, 633
points approaching each other, 627
the annulus, 632–634
the complex plane $\mathbb{C}$, 598
the sphere $\hat{\mathbb{C}}$, 609–610
the torus, 646–651
the upper plane $\bar{\mathbb{H}}$ (bordered), 599
the upper plane $\mathbb{H}$, 599
rotating open string
coherent states, 530
detailed analysis, 140–142
further study, 149
in light-cone gauge, 191
quantum length, 561
quantum states, 526–531

scalar field
action principle, 195–197
degrees of freedom, 199
free equation of motion, 197
free Hamiltonian, 197
light-cone field equation, 200
plane wave solutions, 197–199
quantum theory, 200–206
Schwarz–Christoffel map
closing off the polygon, 628
for four open strings, 618–621
for three open strings, 605–608
in general, 603–605
turning angle, 603
sectors
for open strings between D-branes, 339
for type II superstrings, 322–324
Ramond and Neveu–Schwarz, 312
self-dual radius, 391
Shapiro–Virasoro amplitude, 657
slope parameter α'
defined, 168
related to string length, 170
related to string tension, 169
spacelike
separated events, 16
vector, 21
spatial surface, 100, 121

sphere
  arbitrary dimension, 52
  volume, 54
Standard Model
  first described, 5–6
  gauge group, 457–463
  matter content, 463–472
  string model, 472–479, 492–493
state space
  closed string
    no compactification, 290–294
    with compactification, 388–392
  Lorentz covariant, 577–580
  open string
    between Dp- and Dq-branes, 344
    Dp-brane, 331–333
    parallel Dp-branes, 338–345
    spacefilling D-brane, 262–268
static gauge
  defined, 116
  string action in, 123
string bit, 515–516, 524
string charge
  $\vec{Q}$, 360, 373
  behaving as Maxwell current, 362–363
  carried by electric field, 369–370
  density $\vec{j}^0$, 359–361
string coordinate, 107
string coupling
  and Newton's constant, 294
  change under T-duality, 414
  related to the dilaton, 294–296
string endpoints
  are charged, 368
  covariant analysis of motion, 127
  exercises on free motion, 149
  free motion, 124–125
  motion attached to D-branes, 128
string length, 14
  defined, 170
  related to α', 170
  self-dual radius, 391
string number $\mathcal{N}$, 363
string orientation, 360
string spatial surface, 121
string tension
  effective, 134
  nonrelativistic string, 73
  relativistic string, 111, 119
supergravity, 518
superstrings
  charged D-branes, 371
  closed superstrings, 322–324
  GSO truncation, 320–322
  heterotic, 324, 326, 479
  NS sector, 312–315
  open superstrings, 320–322
  R sector, 315–317
  type I, 324
supersymmetry, 6, 478, 518
symmetry, *see* Lagrangian, symmetry

T-duality
  as torus duality, 376
  closed strings
    as $X_R \to -X_R$, 396
    from spectrum coincidence, 392–394
  effect on string coupling, 412, 414
  open string
    invariance of Hamiltonian, 413
    of DN coordinate, 413
  open strings
    derivation, 402–404
    effect on D-branes, 404
    motivated, 400–402
tachyon
  scattering amplitude, 625
  states in closed string, 291
  states in open string, 266
  tachyon potential, 273, 278
timelike
  separated events, 16
  vector, 21
transverse oscillation
  defined, 73
  frequencies, 77
  wave equation, 75

ultraviolet divergences, 630–631
units
  basic, 13–14
  Heaviside–Lorentz, 45
  natural, 177–178
  Planckian, 60
  SI system, 13–14
unoriented
  closed strings, 304
  open strings, 268, 278, 325
  type I theory, 324

vacuum energy
  associated length scale, 70
  in flux compactification, 486–489, 494
  in the universe, 486
  in toy model, 485, 494
Veneziano amplitude, 625
Virasoro
  algebra
    as a Lie algebra, 276
    consistency of central extension, 276
    subalgebra, 276
    with central extension, 257
    without central extension, 255

descendent, 576–577, 588
null state, 577
operators
covariant closed strings, 580–582
covariant open strings, 571
light-cone closed strings, 286–287
light-cone open strings, 250–259
ordering of $L_0^\perp$, 251–253
primary, 575, 588
transverse modes, 189

Weyl transformation, 586
Wilson line
defined, 408
effect on particle spectrum, 410
holonomy, *see* holonomy
with constant A_x, 409
winding
closed strings, 378–381
defined, 381
number, 362
string charge, 398
Witt algebra, 255
world-line, 16, 90, 107
world-sheet
class of parameterizations, 181
current $\mathcal{P}_\mu$, 159–161
defined, 100
fermion, 309
momentum, 290

Yang–Mills
$U(1)$ theory, 344
$U(3)$ vs. $SU(3)$, 458–462
$U(N)$ gauge fields, 344

zeta-function
analytic continuation, 275
defined, 253

Cosmology and String Theory

Contents

Part I Standard Cosmology

1 The Expanding Universe and the Big Bang 3
1.1 Observations 3
1.2 Newtonian Approximation for the Expansion of Space 5
1.3 Time Evolution of the Scale Factor for a Linear Equation of State 8
1.4 Universe with Several Matter Components 10
1.5 A Quick History of Experimental Observations and Discoveries 11
1.6 The Cosmological Constant 12

2 Relativistic Theory 15
2.1 The FLRW Metric and the Friedmann Equation 15
2.2 Types of Matter and Resulting Expansion of the Universe 19
2.3 Cosmological Evolution Formulas 21

3 The Propagation of Light and Measurements of Distance, Luminosity and Mass 27
3.1 Propagation of Light 27
3.2 Measurements of Distance by Light 29
3.3 Optical Measurements 30
3.4 Stellar Evolution 33
3.5 Measurements of Distance 34
3.6 Mass Determination for Distant Bodies 37

4 Evidence for Dark Matter and the ΛCDM Model 41
4.1 Evidences for Dark Matter 41
4.2 Candidates for Dark Matter 46
4.3 WIMP Dark Matter Candidates 48
4.4 The ΛCDM Model 50

5 The Early Universe and Its Thermal History 53
5.1 Defining the Hot Big Bang 53
5.2 Equilibrium Thermodynamics 55
5.3 Thermodynamics of the Universe 58
5.4 Particle Species in the Radiation Dominated (R.D). Era 61
5.5 Neutrino Decoupling 63

6 Big Bang Nucleosynthesis and Recombination 67
6.1 Recombination Temperature 67
6.2 Big Bang Nucleosynthesis 69
6.3 Dependence of BBN Primordial Abundances on Parameters 72
6.4 Observations and Comparison with BBN 73
6.5 Summary of the Thermal History 74

7 The Cosmic Microwave Background Radiation (CMBR) Anisotropy 77
7.1 The CMBR and Its Anisotropy 77
7.2 Kinematics of the CMBR 79
7.3 Perturbation Theory for CMBR 83
7.4 Effect of Density Perturbations on the CMBR 89

8 Problems to Be Solved by Inflation and How They Are Solved in Inflationary Models 95
8.1 Problems with Standard Cosmology Before Inflation 95
8.2 The Paradigm of Inflation 99
8.3 Inflationary Solutions to Cosmological Problems 101
8.4 Inflation with a Scalar Field 106

9 Slow-Roll Inflation 111
9.1 Slow-Roll Analysis 112
9.2 General (Non-slow Roll) Formulas 115
9.3 Slow Roll Conditions 117
9.4 Exact Solution for Inflation 118
9.5 Variants of Inflation 119

10 Reheating and Baryogenesis 123
10.1 Standard Reheating 123
10.2 New Reheating and Preheating 127
10.3 Entropy Production and the Number of e-Folds 129
10.4 Baryogenesis 130
10.5 Relics 131

11 Fluctuations in Inflation and Matching with Experimental Data 133
11.1 Scales During Inflation 133
11.2 Scalar Fluctuations During Inflation 134
11.3 Quantization of Scalar Perturbations 136
11.4 Scalar-Gravity Fluctuations 138
11.5 Power Spectrum and Contact with Experimental Data 141
11.6 Primordial Tensor Perturbations 143
11.7 Constraints on Inflationary Models 145

Part II Elements of String Theory

12 Extra Dimensions and Kaluza–Klein 151
12.1 KK Metrics 152
12.2 Fields with Spin 154
12.3 The Original Kaluza–Klein Theory 156
12.4 Generalization: The n-Torus $T^n = (S^1)^{\times n}$ 156

13 Electromagnetism and Gravity in Various Dimensions. Consistent Truncations 159
13.1 Review of Electromagnetism and Gravity 159
13.2 Interlude: Spheres in Higher Dimensions 160
13.3 Gauss's Law for Electromagnetism in d Dimensions 161
13.4 Gauss's Law for Gravity in d Dimensions 162
13.5 Consistent Truncations 164

14 $\mathcal{N} = 1$ Supergravity in 4 Dimensions 169
14.1 Rigid Supersymmetry 169
14.2 Supergravity 172
14.3 Vielbein-Spin Connection Formulation of General Relativity 174
14.4 $\mathcal{N} = 1$ Supergravity in 4 Dimensions 176

15 KK Compactification of Supergravity Models 181
15.1 KK Reduction 182
15.2 $\mathcal{N} = 1$ Supergravity Coupled to Matter 183
15.3 Supergravity Compactifications 186
15.4 11 Dimensional Supergravity 187

16 The Relativistic Point Particle 191
16.1 Relativistic Second Order Action 191
16.2 Quantum Field Theory in the Worldline Formalism 194
16.3 First Order Particle Action 195
16.4 Quantization of the Particle in the Light-Cone Gauge 197

17 Relativistic Strings 201
17.1 String Worldsheet and Tension 201
17.2 Induced Metric and the Nambu–Goto Action 202
17.3 The First Order Polyakov Action 204
17.4 Equations of Motion, Boundary Conditions and Constraints 205
17.5 Coupling of String to Background Fields 209

18 Light-Cone Gauge Strings and Quantization 213
18.1 Light-Cone Gauge 213
18.2 String Mode Expansions 215
18.3 Constraints and Hamiltonian 216
18.4 Quantization of String Modes 219
18.5 Light-Cone Gauge Quantization 220

19 D-Branes and Gauge Fields 225
19.1 The D-Brane Action 226
19.2 Chan–Paton Factors and Several D-Branes 231
19.3 Quantizing Open Strings on Dp-Branes 231

20 Electromagnetic Fields on D-Branes and $\mathcal{N} = 4$ SYM. T-Duality of Closed Strings 237
20.1 Supersymmetric String Theory 237
20.2 D3-Brane Action and $\mathcal{N}= 4$ SYM 238
20.3 Nonlinear Born–Infeld Action 240
20.4 Closed Strings on Compact Spaces 241
20.5 T-Duality of Closed Strings 244

21 T-Duality of Open Strings. M-Theory and the Duality Web 249
21.1 T-Duality of Open Strings 249
21.2 T-Duality with Chan–Paton Factors: Several D-Branes 251
21.3 Supergravity Actions 253
21.4 M-Theory 257
21.5 The String Duality Web 259

22 String Theory and Particle Physics 263
22.1 Dimensional Reduction on Compact Spaces 263
22.2 Calabi–Yau Compactification of String Theories 266
22.3 Intersecting D-Branes: Braneworld 268
22.4 Moduli Stabilization 269

23 Holography and the AdS/CFT Correspondence 271
23.1 AdS Space and Its Holography 271
23.2 Conformal Field Theories 274
23.3 AdS/CFT Motivation (Heuristic Derivation) 276

23.4 AdS/CFT Definition and Limits . . . 279
23.5 Operator/State Map and GKPW Construction . . . 280
23.6 Generalization to Non-conformal Theories: Gauge/Gravity Duality . . . 283

Part III Introduction to String Cosmology

24 Problems of String Inflation . . . 289
24.1 Moduli of Type IIB String Compactifications on CY_3 . . . 290
24.2 Approximations for String Calculations of the Potential . . . 293
24.3 Eta Problem in String Theory . . . 294
24.4 Energy Scales for Most Controlled Approximations . . . 295
24.5 Large Field Inflation . . . 297

25 Problems of the Supergravity Approximation to String Inflation . . . 299
25.1 Supergravity Plus Chiral Multiplet . . . 299
25.2 D-Term Inflation . . . 301
25.3 Field Redefinitions . . . 303
25.4 Example with $\epsilon, \eta \ll 1$ . . . 303
25.5 Special Embedding of Inflationary Potentials in Supergravity . . . 306

26 Brane (-Antibrane) Inflation . . . 311
26.1 Brane Probe Inflation . . . 311
26.2 Brane-Antibrane Inflation: Flat Space Approximation . . . 314
26.3 Brane-Antibrane Inflation in Compact Space . . . 317

27 Braneworld Cosmology and the Israel Junction Conditions . . . 321
27.1 Braneworld Cosmology in the Brane Formalism . . . 322
27.2 Embedding Theory . . . 323
27.3 Braneworld and Cosmology . . . 325
27.4 Israel Junction Conditions . . . 326
27.5 FLRW Cosmology . . . 328

28 The KKLT Scenario for de Sitter Backgrounds in String Theory . . . 331
28.1 Supersymmetric Vacua Are AdS . . . 331
28.2 General Set-up for KKLT . . . 333
28.3 Moduli Stabilization in KKLT . . . 335
28.4 Compactification . . . 337
28.5 de Sitter Vacua . . . 338

29 The KKLMMT Scenario for Inflation and Generalizations . . . 341
29.1 Simplified Model for KKLMMT . . . 342
29.2 String Theory Example for KKLMMT Construction . . . 345

29.3 Volume Stabilization and Problems . . . 346
29.4 KKLMMT Generalization with Magnetic Flux . . . 348
29.5 Type IIA Model Corresponding to KKLMMT . . . 349

30 The Ekpyrotic Scenario . . . 353
30.1 The Ekpyrotic Phase . . . 354
30.2 Kinetic Phase . . . 356
30.3 String Theory Model . . . 359
30.4 Solving an Issue Related to the Null Energy Condition . . . 361
30.5 Fluctuations in Ekpyrotic Models . . . 364

31 The Cyclic and New Ekpyrotic Scenarios . . . 367
31.1 The Cyclic Model . . . 367
31.2 The New Ekpyrotic Cosmology . . . 372

32 String Gas Cosmology: Basics and Brandenberger-Vafa Scenario . . . 379
32.1 Hagedorn Temperature . . . 379
32.2 3 Large Spatial Dimensions . . . 381
32.3 String Equations of Motion . . . 382
32.4 String Gas Cosmological Model . . . 383
32.5 Cosmological Early Time Evolution . . . 387
32.6 Moduli Stabilization . . . 388

33 String Gas and Brane Gas Developments . . . 391
33.1 Brane Gas . . . 391
33.2 Cosmological Perturbations Generalities . . . 393
33.3 Correlators from Thermodynamics . . . 395
33.4 Thermodynamics for String Gas . . . 397
33.5 String Gas Power Spectra and Tilts . . . 399

34 Chameleon Scalars and String Theory . . . 405
34.1 Chameleon (and Symmetron) Models . . . 405
34.2 Thin-Shell Effect . . . 408
34.3 String Theory Embedding . . . 409
34.4 Experimental Constraints . . . 411

35 Axion Inflation and Axion Monodromy from String Theory . . . 415
35.1 Axions in Field Theory . . . 415
35.2 Axions in String Theory . . . 417
35.3 "Racetrack" with Two Axions . . . 420
35.4 N-Flation . . . 422
35.5 Axion Monodromy Inflation . . . 423
35.6 Dante's Inferno . . . 426
35.7 Phenomenology of Axion Monodromy Inflation . . . 427

36 Fuzzy Dark Matter from String Theory 431
36.1 Fuzzy Dark Matter from String Theory 431
36.2 Axion Dynamics in the Expanding Universe 432
36.3 Fuzzy Dark Matter as a Nonrelativistic Quantum System 433
36.4 Superfluid Picture 434
36.5 Schrödinger–Poisson Equation and Spherical Soliton Solution 435
36.6 Observational Consequences 436

37 Holographic Cosmology 439
37.1 Generic Holographic Cosmology Map 439
37.2 Domain Wall/Cosmology Correspondence 441
37.3 Holographic Calculation 443
37.4 Phenomenological QFT Calculation 444
37.5 Time Dependent Field Theory Model: Top Down Holography 447
37.6 Spectrum of Fluctuations Going Through Singularity, from Holography 449

References 453

Index 457

Introduction

This book describes the intersection of string theory with cosmology or string cosmology. It is intended for people, either graduate students or researchers, with only a superficial understanding of one, or both of the sides (cosmology or string theory). The necessary background for the book is thus a thorough understanding of general relativity and quantum field theory, and some background in cosmology and string theory, but not too much.

In Part I of the book, I introduce regular cosmology, with the emphasis on the issues that will be explored using string theory, but trying to introduce all relevant concepts and most relevant calculations. Part II introduces the basic elements of string theory, with added emphasis on issues that will be used for cosmology. Besides supergravity and KK compactification, the basics of relativistic strings and their quantization, I will introduce D-branes, extended soliton-like objects in string theory, and T-duality, relevant in some cosmological models, as well as the general picture of supersymmetric string theory, their strong coupling version called M-theory, and their interconnection in the duality web. Finally, I will describe how one obtains particle physics from string theory, and the connection of string theory to strongly coupled field theories called the AdS/CFT correspondence, that can be used in a certain comological model of "holographic cosmology."

Part III of the book contains string cosmology per se. String theory has not managed yet to obtain a fully consistent model of string inflation; there is always a final step missing in those attempted constructions. I will thus describe the problems of constructing a string inflationary model as well as the problems with the supergravity approximation to string inflation. I describe several (related) attempts within the general category of string inflation, brane-antibrane inflation, the KKLT scenario for de Sitter backgrounds, and the KKLMMT scenario for string inflation and its generalizations, as well as the needed technical knowledge (for the previous constructions) of general braneworld cosmology and the Israel junction conditions. Next, I describe the competing models of ekpyrotic, new ekpyrotic and cyclic models, as well as the string gas and brane gas ones. Chameleon scalars, a type of

scalars which are very light on the largest cosmological scales, are obtained from string theory as well as fuzzy dark matter (a kind of axionic scalar model). Finally, I treat the more modern and promising models of axion inflation and axion monodromy from string theory and of holographic cosmology (which uses the AdS/CFT correspondence).

Part I
Standard Cosmology

Chapter 1
The Expanding Universe and the Big Bang

In part I of the book, I will quickly describe the relevant parts of standard (non-stringy) cosmology.

In this first chapter, we will begin with a study of the basics of cosmology in a non-relativistic (non-general relativity) setting. Of course, the Universe is expanding, and the proper description of that is in general relativity, but we will see that with a simple non-relativistic analysis, supplemented with a few nontrivial points, we can go quite far in the description of cosmology. This can be thought of as just a useful trick in order to get some simpler physical insight. In the next chapter, we will move on to the relativistic theory.

After describing some observational facts, we will describe the cosmology in non-relativistic setting, and we will see that the equation of state of "matter" in the Universe dictates the time evolution of the scale factor of the Universe. After that, we will finally describe the most (general) relativistic component of "matter" in the Universe, the cosmological constant, which has a negative pressure, and leads to acceleration, instead of attraction and slowdown, of the expansion.

1.1 Observations

We will start with a few basic observational facts about the Universe.

Fact 1. The most important is the observation of the Hubble expansion, or *Hubble law*.

One defines the *redshift* as the relative change in frequency of light due to the (special relativistic) Doppler effect, which relates the frequency at emission ν_0 with the observed ("redshifted") frequency ν

$$\nu_0 = \nu \left(1 + \frac{v}{c}\right) , \tag{1.1}$$

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_1

so that the redshift is

$$z \equiv \frac{\nu_0 - \nu}{\nu} = \frac{v}{c}. \tag{1.2}$$

Since $\nu = c/\lambda$, we can also write it in terms of the wavelength at emission λ_0 and the observed wavelength λ,

$$z = \frac{\lambda - \lambda_0}{\lambda_0}. \tag{1.3}$$

A measure of the distance r is the *apparent visual magnitude* (more precisely, the d_L defined later is). One measures the light coming from a source, and knowing its brightness, we can see how far away from us it is. Then in the graph of the redshift z with the apparent visual magnitude, standing in for r, we see a linear relation for astrophysics objects, namely stars, galaxies, etc. This is known as the *Hubble law*, and it is

$$(c)z = v = Hr. \tag{1.4}$$

More precisely, this is correct in the non-relativistic approximation. In general, the relation is not linear, and $H = H(t)$. Note that here $v = \dot{r}$.

Finally, we should say that the Hubble law means that it is space itself that expands, and stars and galaxies only go along with the expansion, since there can be no mechanism that all stars and galaxies move away from us according to the Hubble law, and yet space is static.

Fact 2. The other important observation is the one of *isotropy* of the Universe. To infer it, we will preview the fact (to be explained in detail later in the book) that there exists a background radiation permeating all of the Universe, the Cosmic Microwave Background Radiation (CMBR), which is a remnant of the Big Bang, and this radiation has travelled to us since then. The CMBR spectrum is isotropic to an excellent precision, which means that the Universe (the initial Big Bang point, as well as everything in the path of the light coming from the Big Bang, that is to say, the whole of the observable Universe) is isotropic.

There is one detail about the CMBR that necessitates a relativistic treatment in order to understand properly. The CMBR has a dipole, i.e., there is a preferred direction in space. But this is easily understood as being due to the fact that the Earth is moving at a constant velocity, in the direction of this dipole, with respect to the CMBR. This is not clear within special relativity, where there shouldn't be a preferred frame for radiation. But within general relativity, in the expanding Universe metric $ds^2 = -c^2dt^2 + a^2(t)d\vec{x}^2$, there is a preferred frame, defined by this metric, where t is the "cosmological time". It is in this frame that the CMBR is defined, and the Earth is moving with respect to it.

Fact 3. Finally, we can observe the *homogeneity* of the Universe. Of course, on small scales, the Universe is not homogenous: we have stars and galaxies, either of which is surrounded by voids until the next structure. But on even larger scales, where galaxies and clusters of galaxies are considered as point-like, the Universe is homogenous. We can explicitly see this, from the distribution of galaxies and clusters

of galaxies, which is found by measuring the direction, distance and mass of these structures. But moreover, while this method simply uses light to detect matter, there are other forms of matter that do not appear in light, called dark matter, about which we will speak later on in the book. To take that into consideration, one can measure the *peculiar velocity* fields of stars and galaxies. As we said, on the average, stars and galaxies move away from us according to the Hubble law (due to the expansion of space), but still, relative to this expansion of space, the objects can have extra random velocities, called peculiar velocities. In gravitational theory, the velocities of these many objects are influenced by the gravitational fields present, generated by the local mass distribution. Therefore, by measuring the peculiar velocities, in effect we can extract information about the mass distribution, that includes any matter not visible via light, i.e., dark matter. The final result is the same: on the very largest scales (of galaxies and clusters of galaxies), the Universe is homogenous to a very high degree.

1.2 Newtonian Approximation for the Expansion of Space

In order to find the equations that describe the expansion of space, we must use gravitational equations and the conservation of energy. The simplest model, that we will use in this chapter, is to use the Newtonian approximation for gravity. I should stress here that, strictly speaking, Newtonian gravity is not applicable to the expanding Universe, which is described by general relativity. However, the simplest model, Newtonian gravity, can take us very far, and by including the relativistic effect of pressure on the energy, can give us the correct equations for the expanding Universe. The method presented here is a rather known trick, whose origins I don't know, but which I have first seen in lectures by Igor Novikov some 25 years ago. The goal, besides ease of introduction, is to see that, despite appearances, we only need minimal input from general relativity, otherwise we can derive most things without it.

Assuming spherical symmetry (from isotropy) and uniformity, we can consider a sphere of matter of constant energy density ρ, and of radius r. Because of the expansion of space through the Hubble law, we have

$$\dot{r} = v = Hr. \tag{1.5}$$

Then the gravitational Gauss's law equates the total mass in the sphere (times $4\pi G_N$) with the flux of the gravitational field through the surface of the sphere, i.e.,

$$\rho \times \frac{4\pi}{3} r^3 \times (4\pi G_N) = g \times 4\pi r^2. \tag{1.6}$$

Moreover, g equals (since $m_i = m_g$, the inertial and gravitational masses are the same) the acceleration at the surface of the sphere, i.e.,

$$\ddot{r} = g = -\frac{4\pi G_N}{3}\rho r. \tag{1.7}$$

This is in fact the same as the general relativity result for pressureless matter! The only effect of general relativity is to imply that pressure has energy density as well, resulting in the shift

$$\rho \to \rho + \frac{3P}{c^2}\ , \tag{1.8}$$

to finally obtain the general relativity relation (renaming $r \to R$, the scale factor of the Universe)

$$\ddot{R} = -\frac{4\pi G_N}{3}\left(\rho + \frac{3P}{c^2}\right) R. \tag{1.9}$$

The Hubble law in general has

$$H = H(t) \equiv \frac{1}{R(t)}\frac{dR(t)}{dt}\ , \tag{1.10}$$

so the Hubble "constant" is not really a constant, but varies in time (it is a constant only in space). Moreover, as a function of the redshift z (the observational quantity), H also varies, since we measure everything by the light that comes to us, and something coming from redshift z corresponds (due to the constancy of the speed of light) with something that comes from a certain distance d, and thus a certain time t away.

We can also obtain another law from the Newtonian approximation. We can obtain it more rigorously, but first we will use a not quite correct shortcut. Neglect the pressure, $P \simeq 0$, and consider the conservation of energy in its rawest form, the sum of the kinetic and potential energies be a constant for a test mass m. The kinetic energy is $m\dot{R}^2/2$, and a very naive estimate for the potential energy, m times the potential, is m times the mass of the sphere, $\rho(4\pi R^3/3)$, times $1/R$, obtaining

$$\frac{m}{2}\left(\frac{dR}{dt}\right)^2 - m\left(\frac{4\pi}{3}R^3\rho\right)\frac{1}{R} = \text{const.} \tag{1.11}$$

But of course, this derivation is not quite correct, since to obtain the potential we would need to actually integrate the gravitational acceleration, $\int g(R)dR$, and then we would get a factor of 2 too much with respect to the correct result.

The correct way to derive the above result is to use the relation previously derived, (1.9), together with *local* energy conservation. Note that from now on we will put $c = 1$ as usual in theoretical physics. The latter amounts to writing the energy variation dU as $d(\rho V) = Vd\rho + \rho dV$, and then equating with the work done by pressure, $-PdV$, to obtain

$$\dot{\rho} = -(\rho + P)\frac{\dot{V}}{V} = -3(\rho + P)\frac{\dot{R}}{R}. \tag{1.12}$$

Eliminating P between this equation and (1.9), we obtain

$$\ddot{R} = +\frac{8\pi G_N}{3}\rho R + \frac{4\pi G_N}{3}\dot{\rho}\frac{R^2}{\dot{R}}. \tag{1.13}$$

Multiplying the latter by $\dot{R}$, we obtain

$$\frac{d}{dt}\left(\frac{\dot{R}^2}{2}\right) = \frac{d}{dt}\left(\frac{4\pi G_N}{3}\rho R^2\right) , \tag{1.14}$$

which integrates to

$$\frac{\dot{R}^2}{2} - \frac{4\pi G_N}{3}\rho R^2 = \frac{k}{2} = \text{const.} \tag{1.15}$$

We see that this is the same equation as from the naive (and wrong) derivation, but now it is correct. This equation is known as the *Friedmann equation*. Moreover, we can define the constant through the value of the left hand side at current time, denoted by a zero index, obtaining

$$\left(\frac{dR}{dt}\right)^2 = \frac{8\pi G_N}{3}\rho R^2 - \frac{8\pi G_N}{3}R_0^2\left(\rho_0 - \frac{3H_0^2}{8\pi G_N}\right) , \tag{1.16}$$

where the second term on the right hand side is the constant k.

We now define the *critical energy density* at current time

$$\frac{3H_0^2}{8\pi G_N} \equiv \rho_{c,0} , \tag{1.17}$$

and the ratio of the energy density at current time with the critical one,

$$\Omega_0 \equiv \frac{\rho_0}{\rho_{c,0}} = \frac{8\pi G_N}{3}\frac{\rho_0}{H_0^2}. \tag{1.18}$$

We can define the same quantities at arbitrary times, the critical energy density

$$\frac{3H^2}{8\pi G_N} \equiv \rho_c , \tag{1.19}$$

and the ratio of the energy density to the critical one,

$$\Omega \equiv \frac{\rho}{\rho_c} = \frac{8\pi G_N}{3}\frac{\rho}{H^2}. \tag{1.20}$$

Now the information that comes from general relativity and will be explained in the next chapter, is that $\Omega > 1$ implies that k is not an arbitrary constant, but rather

it is $k = +1$, and the Universe is (spatially) *closed*, i.e., it something like a sphere, with a finite total volume. If on the other hand $\Omega < 1$, this means that $k = -1$, and the Universe is (spatially) open, a hyperbolic space of infinite total volume. But in fact, experimentally we are almost certainly in the case $\Omega = 1$, which implies $k = 0$, and the Universe is flat, a parabolic space of infinite total volume.

Finally, that means that we can rewrite the Friedmann equation as

$$\frac{k}{H^2R^2} = \frac{\rho}{3H^2/(8\pi G_N)} - 1 = \Omega - 1. \tag{1.21}$$

If H quantifies the velocity of the expansion, we need to define a parameter that describes also its acceleration. Or rather, we define the *deceleration parameter*

$$q \equiv \frac{1}{H^2}\left(-\frac{1}{R}\frac{d^2R}{dt^2}\right) \tag{1.22}$$

Then from the equation for $\ddot{R}$ (1.13), we obtain

$$q = -\frac{\ddot{R}}{RH^2} = -\Omega\left(1 + \frac{1}{2}\frac{\dot{\rho}/\rho}{\dot{R}/R}\right). \tag{1.23}$$

The parameter is defined as a deceleration parameter (as opposed to an acceleration parameter), since historically it was thought that the Universe is decelerated, though as we will see, now we know it currently is accelerated.

For pure pressureless matter (dust), $P = 0$, $\dot{\rho}/\rho = -3\dot{R}/R$, which gives

$$q = \frac{\Omega}{2}. \tag{1.24}$$

Experimentally, the Hubble parameter today H_0 is defined in terms of a dimensionless one h_0 as

$$H_0 \equiv h_0 \times 100\frac{\text{Km/s}}{\text{Mpc}}. \tag{1.25}$$

1.3 Time Evolution of the Scale Factor for a Linear Equation of State

In cosmology, one usually assumes a linear equation of state,

$$P = w\rho. \tag{1.26}$$

This is of course true for dust ($P = 0$), and for radiation $P = \rho/3$, i.e., $w = 1/3$, but is not necessarily true for any kind of matter. For instance, for a scalar field in

a potential, we can have a varying w (nonlinear $P(\rho)$), and for QCD in the phase transition, we also have a nonlinear $P(\rho)$. Nevertheless, we will assume constant w in the following.

Then the continuity equation is

$$\frac{\dot{\rho}}{\rho} = -3(1+w)\frac{\dot{R}}{R}\,, \tag{1.27}$$

and we see that we don't need to eliminate P anymore, so we can integrate it directly,

$$\rho(R) \propto R^{-3(1+w)}. \tag{1.28}$$

If $\Omega = 1$, like it seems to be experimentally favoured, or if we are looking at $R \to \infty$, from (1.21) we obtain

$$\frac{\dot{R}^2}{R^2} = H^2 = \frac{8\pi G_N}{3}\rho \propto R^{-3(1+w)}\,, \tag{1.29}$$

which we can integrate to give

$$R \propto t^{\frac{2}{3(1+w)}}. \tag{1.30}$$

From this, we obtain

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{4}{9(1+w)^2 t^2}\,, \tag{1.31}$$

to be compared with $8\pi G_N \rho/3$, so that we have

$$\rho = \frac{1}{6\pi G_N t^2 (1+w)^2}. \tag{1.32}$$

From $R(t)$, written (by comparing with the relation today) as

$$R = R_0 (t/t_0)^{\frac{2}{3(1+w)}}\,, \tag{1.33}$$

we get

$$H_0 = \frac{2}{3(1+w)t_0} \Rightarrow t_0 = \frac{2}{3(1+w)} H_0^{-1}. \tag{1.34}$$

Thus the age of the Universe, at least in this simple model where a single w dominates the whole evolution of the Universe, is given roughly by H_0^{-1}.

1.4 Universe with Several Matter Components

Consider several energy components with different equations of state $P_i = w_i \rho_i$. Then each component has its own continuity equation (the energy is conserved for each component, since for each we have $dU_i = d(\rho_i V) = -P_i dV$)

$$\frac{\dot{\rho}_i}{\rho_i} = -3(1 + w_i)\frac{\dot{R}}{R} , \tag{1.35}$$

which integrates to

$$\rho_i(R) \propto R^{-3(1+w_i)}. \tag{1.36}$$

Thus each component decays with R in a different way.

On the other hand, if $\Omega = 1$ or $R \to \infty$, we have the Friedmann equation in terms of the total energy, i.e.,

$$\left(\frac{\dot{R}}{R}\right)^2 = H^2 = \frac{8\pi G_N}{3}\sum_i \rho_i. \tag{1.37}$$

If one of them, w_{dom}, dominates the energy density, then $\rho \simeq \rho_{dom}$, so we obtain

$$R \propto t^{\frac{2}{3(1+w_{dom})}} , \tag{1.38}$$

which leads to the scaling of the total energy density as

$$\rho = \frac{1}{6\pi G_N t^2 (1 + w_{dom})^2}. \tag{1.39}$$

But since $\rho_i(R)$ scale with R in a different way, their relative proportion changes, since

$$\rho_i(R) \propto R^{-3(1+w_i)} \Rightarrow \frac{\rho_i}{\rho_j} \propto R^{3(w_j - w_i)}. \tag{1.40}$$

That means that the component with the lowest w_i dominates at late times, and the one with the highest w_i dominates at early times.

Several components in the Friedmann equation for $\ddot{R}$ gives

$$\frac{\ddot{R}}{R} = -\frac{4\pi G_N}{3}\sum_i (\rho_i + 3P_i) = -\frac{4\pi G_N}{3}\sum_i \rho_i(1 + 3w_i) , \tag{1.41}$$

and defining the relative Ω's,

$$\Omega_i = \frac{8\pi G_N \rho_i}{3H^2} , \tag{1.42}$$

we can write the Friedmann equation as the equation for the deceleration parameter

$$q \equiv -\frac{\ddot{R}}{RH^2} = \sum_i \frac{\Omega_i}{2}(1+3w_i). \tag{1.43}$$

1.5 A Quick History of Experimental Observations and Discoveries

At this point, we pause and give a quick timeline of the relevant discoveries for the standard cosmology described so far.

–In 1914, Slipher measures nebular velocities v, and finds that they are much larger than stellar velocities.

–In 1917, the same Slipher measures the average velocity as $\langle v \rangle = 500\,\mathrm{km/s}$. He is the first to observe the galactic redshift. Notice that Einstein's general relativity was discovered in 1915, in between the two discoveries.

–In 1917: de Sitter writes his solution with repulsive forces proportional to the distance giving an expanding Universe. It is the first astrophysical implication of Einstein's general relativity. Einstein however doesn't like it, and invents "Einstein's static Universe" to counter it.

–In 1922 and 1924, Friedmann writes his equations for the expanding Universe.

–In 1923, Hermann Weyl describes galaxies in de Sitter's Universe.

–In 1927, Lemaître writes his theory.

–In 1928, Robertson writes his, and later Walker. Thus finally, we have the FLRW Universe, to be described in the next chapter.

–On 15 March 1929, E. Hubble finds his law for the expanding Universe. He measures 24 nebulae with known distances (we will see later how we can measure distance, which is nontrivial), and measuring also the redshift, finds $H \simeq 500\,\mathrm{Km/s/Mpc}$. But note that now we know that $H < 100\,\mathrm{Km/s/Mpc}$, so in reality his measurement was way off, and yet he managed to establish the correct expansion law.

Note that by inverting the value of H_0 of about $100\,\mathrm{Km/s/Mpc}$, we obtain the order of magnitude of the age of the Universe,

$$t_0 \sim H_0^{-1} \sim 10^{10} yrs = 10\ billion\ years. \tag{1.44}$$

This is so, since, if the Universe always expanded with the same H_0 (which we know not to be the case, in fact $H = H(t)$, but the variation is only polynomial), at $T_0 = 1/H_0$ in the past all points were situated at the same place.

This is roughly correct, since the exact age of the Universe measured now is about 13.8 billion years.

1.6 The Cosmological Constant

There is a relativistic "matter" component (see the next chapter) called the *cosmological constant* that has $w = -1$, which means that it has the equation of state

$$P = -\rho. \tag{1.45}$$

It can only be understood in relativistic theory, since we have negative pressure. In fact, there are so-called energy conditions that are believed to hold for any quantum matter, one of them being $P + \rho \geq 0$, so the cosmological constant saturates this inequality. In fact, usually one considers $w \in [-1, 1]$, but only the lower value is an actual bound, the upper value is not: we could consider $w \gg 1$, as we will in part III of the book.

Since $w = -1$ is the minimum value possible for w, in the end it will dominate the evolution of the Universe. This is what happens in the current age, when Λ just begins to dominate. It is also what happened during inflation, when there was an effective cosmological constant.

For the pure Λ Universe (cosmological constant dominating), from the continuity equation with $\rho + P = 0$ we obtain

$$\frac{\dot{\rho}}{\rho} = 0 \Rightarrow \rho = \text{const.} \tag{1.46}$$

So indeed the cosmological constant is an energy density that is not only constant in space, but also in time, making it a true constant.

Note that this is somewhat counterintuitive, since it means that despite the fact that the Universe is expanding, the energy density remains constant. That implies that as a sphere expands, energy is created (since the volume increases, but the energy density remains constant). But that is easy to explain, due to the negative pressure that exerts positive work. It is just counterintuitive, since in usual experience, pressure is positive, so exerts negative work, and the energy density drops as a volume expands.

If $\rho =$constant, the Friedmann equation gives

$$H^2 = \frac{8\pi G_N}{3}\rho = \text{const.} \Rightarrow R = R_0 e^{Ht}\ , \tag{1.47}$$

i.e., exponential expansion.

Next, consider a more physical case, with both cosmological constant (which we find experimentally that exists, at the very least as an approximation, in the real Universe), and matter, which is obviously present. This describes well the current age of the Universe. Then from (1.43), we obtain

$$q = \frac{\Omega_{\text{matter}}}{2} - \Omega_\Lambda. \tag{1.48}$$

Notice the sign for Ω_Λ, and since it dominates, we have *acceleration* of the Universe.

In fact, as seen from (1.43), we only need $w \leq -1/3$ to obtain acceleration.

Observing the Cosmological Constant

Observationally, the measurements that decides the existence of the cosmological constant are related to type I_a supernovae. It is not important what they are (though we will describe them later on in the book), but the important point for us is that they are (statistically speaking) standard candles. That is, we know they all have roughly the same brightness, so by measuring the flux that reaches us on Earth, we can calculate the distance. More precisely, we calculate a $d_P(z)$ that will be defined more precisely shortly.

As we said, the Hubble law is only approximate, and more precisely it is a local (in time, thus space) relation. Instead of $z = \dot{R} = HR$, we have

$$c\, dz = -H\, dr. \tag{1.49}$$

Moreover, $H = H(z)$ as we said, and r is the distance travelled by light, divided by the normalized expanding scale, R/R_0.

The measured distance is the integral of dr,

$$d_P = \int dr = c \int \frac{dz}{H(z)}. \tag{1.50}$$

But H depends on z implicitly, via the dependence on t, related to z via the propagation of light, namely (using $dz = -H\, dr/c = -H dt/(R/R_0)$)

$$\frac{dH}{dz} = \frac{dt}{dz}\frac{dH}{dt} = -\frac{R/R_0}{H}\frac{d}{dt}\left(\frac{\dot{R}}{R}\right) = -\frac{R/R_0}{H}\left(\frac{\ddot{R}}{R} - \frac{\dot{R}^2}{R^2}\right) = +H\frac{R}{R_0}(q+1). \tag{1.51}$$

This implies the expansion

$$\frac{1}{H(z)} \simeq \frac{1}{H_0} - \frac{1}{H_0^2}\left(\frac{dH}{dz}\right)\Big|_0 z. \tag{1.52}$$

Finally, substituting in $d_P = c\int dz/H(z)$, we get

$$d_P \simeq \frac{c}{H_0}\left(z - (1+q)\frac{z^2}{2}\right). \tag{1.53}$$

By measuring the curvature of $d_P(z)$ (deviation from a straight line), we find q. This gives a rough linear relation between Ω_Λ and Ω_m. Experimentally, we find $q < 0$, so we have an accelerated Universe, with $q < 0$, so $w_{total} < -1/3$. More precisely, combining the linear relation above with another linear relation, coming from the fact that CMBR measurements (to be described later) imply that $\Omega = \Omega_m + \Omega_\Lambda = 1$

(flat Universe) is favored, we obtain the current result at the intersection of the two lines, with $\Omega_m \simeq 0.3$ and $\Omega_\Lambda \simeq 0.7$.[1]

Important Concepts to Remember

- The Universe expands according to the Hubble law, $v = Hr$. In general, the Hubble law is $\dot{R}/R = H(t)$.
- The Universe is homogenous and isotropic.
- In the Newtonian approximation, the Universe obeys the Friedmann equation $\dot{R}^2/2 - 8\pi G\rho R^2/3 = k/2 =$ constant and the continuity equation $\dot{\rho} + 3(\rho + P)\dot{R}/R = 0$.
- We can define the ratio Ω of the energy density to the critical energy density $\rho_{\rm cr} = 3H^2/8\pi G_N$, so that the Friedmann equation is $\Omega - 1 = k/H^2R^2$.
- The deceleration parameter $q = -\ddot{R}/RH^2$ for the pressureless matter ($P = 0$) case is $q = \Omega/2$.
- Defining the equation of state $P = w\rho$, we have $\rho(R) \sim R^{-3(1+w)}$ and, if it dominates the energy density, $R \propto t^{2/(3(1+w))}$, so that $\rho \propto 1/t^2$.
- For several components, $\rho_i \propto R^{-3(1+w_i)}$ and $q = \sum_i \Omega(1 + 3w_i)/2$.
- The cosmological constant has $P = -\rho$, $\rho =$ constant in time, and if it dominates gives exponential expansion, $R = R_0 e^{Ht}$, with $H^2 = 8\pi G_N \rho/3$.
- The distance travelled by light is $d_P = \int dr = c \int dz/H(r)$.

Further reading: Most of the basic cosmology can be found in [1]. We can also see the books [2, 3].

Exercises

(1) Calculate $q(t)$ and $q(t)/\Omega(t)$ for arbitrary equation of state $P = w\rho$ and arbitrary k.
(2) Calculate the *exact* $\rho(t)$ in a system with two components, one with w_1 and $\rho_{0,1}$ at t_0, and one with w_2 and $\rho_{0,2}$ at t_0.
(3) Consider a Universe with a radiation component ($w = +1/3$), plus a cosmological constant. Calculate $q(\Omega_i)$ and the *exact* $\rho(t)$.
(4) Calculate the next term in the expansion for $d_P(z)$ beyond the quadratic one in (1.53).

[1] We should mention here that, strictly speaking, flatness does not imply an Universe that is open, as we will assume in the following. For instance, a torus is flat (has no curvature), yet has a finite volume. But there is no convincing argument for nontrivial topology of the Universe: in any case, the topology would have to be outside the observable Universe (outside the "horizon"), in order not to contradict experiments, in which case one would need indirect measurements to test for topology. Since there is no convincing way to test for this topology, I will not consider it in the following.

Chapter 2
Relativistic Theory

In this chapter, we re-derive, and extend, the analysis of the previous chapter within general relativity. We will re-derive the Friedmann equation from the relativistic treatment, of the FLRW metric, and then describe the various types of matter in detail, and the corresponding evolution of the Universe.

2.1 The FLRW Metric and the Friedmann Equation

Homogeneity and isotropy (experimentally observed, as we mentioned in the previous chapter) means first that we can write the metric in diagonal form, $-f(dt)^2 + dl^2$. But homogeneity means that f must be independent of r, and then we can redefine t to put it in the form

$$ds^2 = -cdt^2 + dl^2. \tag{2.1}$$

For the spatial part of the metric, dl^2, by homogeneity and isotropy, we need to have a factor $d\tilde{x}^2 + d\tilde{y}^2 + d\tilde{z}^2$, multiplied by a function of $r^2 = \tilde{x}^2 + \tilde{y}^2 + \tilde{z}^2$, and multiplied by a scale factor ($a(t)$ is the radius of curvature and $C = k/a^2$ is the curvature; here $k = 0, \pm 1$). This gives a 3 dimensional space of constant curvature (which is what homogeneity and isotropy implies), whose metric in comoving coordinates, i.e., coordinates that move together with the expansion of the space, quantified by $a(t)$, is uniquely given by

$$ds^2 = a^2(t)\frac{d\tilde{x}^2 + d\tilde{y}^2 + d\tilde{z}}{\left[1 + \frac{k}{4}(\tilde{x}^2 + \tilde{y}^2 + \tilde{z}^2)\right]^2}. \tag{2.2}$$

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_2

Defining spherical coordinates

$$\begin{aligned}\tilde{x} &= \tilde{r}\sin\theta\cos\phi \\ \tilde{y} &= \tilde{r}\sin\theta\sin\phi \\ \tilde{z} &= \tilde{r}\cos\theta\ ,\end{aligned} \tag{2.3}$$

we obtain

$$dl^2 = a^2(t)\frac{d\tilde{r}^2 + \tilde{r}^2(d\theta^2 + \sin^2\theta d\phi^2)}{\left[1 + \frac{k}{4}\tilde{r}^2\right]^2}. \tag{2.4}$$

Using the further transformations

$$\begin{aligned}k = +1 &: \frac{\tilde{r}}{1 + \frac{\tilde{r}^2}{4}} = \sin r \\ k = -1 &: \frac{\tilde{r}}{1 - \frac{\tilde{r}^2}{4}} = \sinh r \\ k = 0 &: \tilde{r} = r\ ,\end{aligned} \tag{2.5}$$

we obtain

$$\begin{aligned}dl^2|_{k=0} &= a^2(t)[dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2)] \\ dl^2|_{k=+1} &= a^2(t)[dr^2 + \sin^2 r(d\theta^2 + \sin^2\theta d\phi^2)] \\ dl^2|_{k=-1} &= a^2(t)[dr^2 + \sinh^2 r(d\theta^2 + \sin^2\theta d\phi^2)].\end{aligned} \tag{2.6}$$

In a 2 dimensional analogy, if we had a metric $dl^2|_{k=0} = a^2(t)[dr^2 + r^2 d\theta^2]$, we would say that the radius of the circle parametrized by θ would be

$$l = 2\pi r a. \tag{2.7}$$

In the $k = +1$ case with $r \to \sin r$ in the metric,

$$l = 2\pi \sin r a\ , \tag{2.8}$$

and in the $k = -1$ case with $r \to \sinh r$ in the metric,

$$l = 2\pi \sinh r a. \tag{2.9}$$

Similarly now, the area of the 2-sphere parametrized by θ, ϕ is (in terms of the area 4π of the unit sphere)

$$\begin{aligned}S(k) &= 4\pi a^2 r^2 : \quad k = 0 \\ &= 4\pi a^2 \sin^2 r; \quad k = +1 \\ &= 4\pi a^2 \sinh^2 r; \quad k = -1.\end{aligned} \tag{2.10}$$

Finally, we can transform to Friedmann–Lemaître–Robertson–Walker (FLRW) coordinates, by

$$\begin{aligned} r' &= \sin r; \quad k = +1 \\ &= r; \quad k = 0 \\ &= \sinh r; \quad k = -1 \ , \end{aligned} \tag{2.11}$$

giving the FLRW Universe, with metric

$$dl^2 = a^2(t)\left[\frac{dr'^2}{1-kr'^2} + r'^2(d\theta^2 + \sin^2\theta d\phi^2)\right]. \tag{2.12}$$

We can substitute this metric in the Einstein equations, and find that they reduce to an equation for the acceleration of a,

$$\frac{d^2a}{dt^2} = -\frac{4\pi G_N}{3}a(\rho + 3P) + \frac{\Lambda}{3}a \ , \tag{2.13}$$

and an equation for H, the *Friedmann equation*,

$$H^2 = \left(\frac{da}{adt}\right)^2 = \frac{8\pi G_N}{3}\rho - \frac{k}{a^2}. \tag{2.14}$$

These are the same equations as obtained before, in the Newtonian analysis, except as we said there, the Friedmann equation comes from integrating the equation for the acceleration (2.13) and as such k is an arbitrary integration constant. Here $k = 0, \pm 1$ depending on whether the Universe is flat, closed or open. But as before, we can also go in the opposite direction, namely using (2.14) and the continuity equation, we obtain (2.13). Another difference is that now the continuity equation is obtained from the Einstein equation as well, as we will shortly see.

The Einstein equation is written as

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = 8\pi G_N T_{\mu\nu} - \Lambda g_{\mu\nu} \ , \tag{2.15}$$

where we have added a separate cosmological constant term in it, even though we could write it as part of the energy-momentum tensor $T_{\mu\nu}$. Note that here, and in the rest of the book, I will use the metric signature convention $(-+++)$ ("mostly plus").

For the energy-momentum tensor, most of the time in cosmology one uses the *perfect fluid* ansatz (non-perfect fluid is mostly only used in the cases of perturbations, without homogeneity and isotropy),

$$T^{\mu\nu} = \rho u^\mu u^\nu + P(g^{\mu\nu} + u^\mu u^\nu) \ , \tag{2.16}$$

where

$$u^\mu = \frac{dx^\mu}{d\tau} \tag{2.17}$$

is the comoving 4-velocity of the fluid, and the tensor structure $g^{\mu\nu} + u^\mu u^\nu$ is transverse to the 4-velocity of the fluid,

$$u_\mu(g^{\mu\nu} + u^\mu u^\nu) = u^\mu - u^\mu = 0 \ , \tag{2.18}$$

since $u^\mu u_\mu = -1$ (obvious in the frame where $u^\mu = (1, 0, 0, 0)$). Then in the same frame, with $u^\mu = (1, 0, 0, 0)$, we find that $T^\mu{}_\nu$ is diagonal, and more precisely

$$T^\mu{}_\nu = \text{diag}(-\rho, P, P, P) \ , \tag{2.19}$$

both in flat Minkowski space and in the FLRW Universe.

Adding the cosmological constant corresponds to shifting the energy-momentum tensor by

$$T_{\mu\nu} \to T_{\mu\nu} - \frac{\Lambda}{8\pi G_N} g_{\mu\nu} \ , \tag{2.20}$$

which amounts to the shift

$$\begin{aligned} \rho &\to \rho + \rho_\Lambda \\ P &\to P + P_\Lambda; \end{aligned} \tag{2.21}$$

where

$$\rho_\Lambda = \frac{\Lambda}{8\pi G_N}; \quad p_\Lambda = -\rho_\Lambda. \tag{2.22}$$

Energy conservation in general relativity corresponds to the conservation of the current corresponding to translation symmetry, i.e., the energy-momentum tensor,

$$D_\mu T^\mu{}_\nu = 0 \ , \tag{2.23}$$

which is a tensor relation, thus can describe energy loss or gain. With $T^\mu{}_\nu = \text{diag}(-\rho, P, P, P)$, it reduces to the usual relation, derived last time from $dU = d(\rho V) = -P dV$,

$$\dot\rho = -3(\rho + P)\frac{\dot a}{a}. \tag{2.24}$$

As in the previous chapter, we can rewrite the Friedmann equation as

$$\frac{k}{a^2} = \frac{8\pi G_N}{3}(\rho(t) - \rho_c(t)) + \frac{\Lambda}{3} = \frac{8\pi G_N}{3}(\rho_{\text{tot}}(t) - \rho_c(t)) \ , \tag{2.25}$$

where the critical energy density is

$$\rho_c \equiv \frac{3H^2}{8\pi G_N} \,, \tag{2.26}$$

in the second equality we have considered the cosmological constant as part of the *total* energy density $\rho_{\rm tot}$ as well, and the left hand side, sometimes called

$$\frac{k}{a^2} \equiv C_G \,, \tag{2.27}$$

the *curvature energy density* (and as such, can be included in the total energy density as well). Note that in this form of the Friedmann equation, we see that the sign of $\rho(t) - \rho_c(t)$ is the same as the sign of k, which is fixed for all time. That is, the relation of $\rho < \rho_c$ ($\Omega < 1$) or $\rho > \rho_c$ ($\Omega > 1$), or $\rho = \rho_c$ ($\Omega = 1$) is true *for all time*. Moreover, it says that if $\rho_{\rm tot}(t) \to \rho_c(t)$ at some time, then (if $k \neq 0$, i.e., if $\rho \neq \rho_c$) we have $a \to \infty$.

2.2 Types of Matter and Resulting Expansion of the Universe

We first review the analysis of the case with linear equation of state, $P = w\rho$, from the previous chapter.

We found that

$$\begin{aligned} a(t) &\propto t^{\frac{2}{3(1+w)}} \\ H = \frac{\dot{a}}{a} &= \frac{1}{t} \propto a^{-\frac{3(1+w)}{2}} \\ \rho = \frac{1}{6\pi G_N t^2 (1+w)^2} &\propto a(t)^{-3(1+w)}. \end{aligned} \tag{2.28}$$

- $P = 0$ for *dust* matter (pressureless). If it dominates the expansion of the Universe, i.e., if we are in an era of *matter domination* (M.D.), we have, according to the above formulae,

$$\begin{aligned} a(t) &\propto t^{2/3} \\ H &\propto \frac{1}{a^{3/2}} \\ \rho = \frac{1}{6\pi G_N t^2} &= \frac{1}{a^3(t)}. \end{aligned} \tag{2.29}$$

  The last relation is expected, since the matter in a comoving volume is conserved during the expansion of the Universe, but the volume expands as a^3, so the density drops as $1/a^3$.

- $P = \rho/3$ ($w = 1/3$) for radiation. If it dominates the expansion of the Universe, i.e., if we are in an era of *radiation domination* (R.D.), we have, according to the general formulae,

$$\begin{aligned} a(t) &\propto t^{1/2} \\ H(t) &\propto \frac{1}{a^2} \\ \rho &= \frac{3}{32\pi G_N t^2} \propto \frac{1}{a^4(t)}. \end{aligned} \tag{2.30}$$

  The last relation is again expected. The energy of a photon is $h\nu = hc/\lambda$, and wavelengths λ are comoving, i.e., they expand together with the Universe, which means that the energy within a comoving volume decays as $1/a$, and the energy density as $1/a^4$.
- $P = \rho$ ($w = +1$), which is usually called "stiff matter" gives, according to the general formulae,

$$\begin{aligned} a(t) &\propto t^{1/3} \\ H &\propto \frac{1}{a^3} \\ \rho &= \frac{1}{24\pi G_N t^2} \propto \frac{1}{a^6(t)}. \end{aligned} \tag{2.31}$$

  The stiff matter is not really the maximum w one can have, though one rarely considers bigger ones. However, we will see towards the end of the book an example of $w \gg 1$. The case $w = +1$ is obtained for a (homogenous and isotropic) background scalar field $\phi = \phi(t)$, when the kinetic term dominates over the potential term (so the opposite to the case of inflation, when the potential dominates over the kinetic term), so that the action is approximately

$$S = -\frac{1}{2}\int d^4x\sqrt{-g}(\partial_\mu\phi)^2 \,, \tag{2.32}$$

  and the energy-momentum tensor is

$$T_{\mu\nu} = -\frac{2}{\sqrt{-g}}\frac{\delta S}{\delta g^{\mu\nu}} = \partial_\mu\phi\partial_\nu\phi - \frac{1}{2}g_{\mu\nu}(\partial_\rho\phi)^2 \,, \tag{2.33}$$

  meaning if $\phi = \phi(t)$ only, we obtain

$$T_{00} = \frac{1}{2}\dot{\phi}^2 = T_{ii} \Rightarrow w = +1. \tag{2.34}$$

- Finally, we review the case of cosmological constant, with $P = -\rho$ ($w = -1$). In this case, the general formulae are singular at $w = -1$, so we must consider them separately. As we saw in the previous chapter, we have

$$\begin{aligned} a(t) &\propto e^{Ht} \\ H &= \text{const.} \\ \rho &= \text{const.} \end{aligned} \tag{2.35}$$

The terms with highest w dominate at the beginning, and the one with lowest w at the end. Since we have $P + \rho \geq 0$ by the energy conditions (believed to be satisfied by all quantum matter), the order is as follows: first, stiff matter dominates, then radiation, then matter, then cosmological constant.

In order to find the expansion age of the Universe, t_0, as a function of H_0 (the Hubble constant today) and Ω_0 (the total Ω today), we first write the Friedmann equation as

$$\begin{aligned} &\frac{k}{a^2} = \frac{8\pi G_N}{3}(\rho_{\text{tot}}(t) - \rho_c(t)) = \frac{H^2(t)}{\rho_c(t)}(\rho_{\text{tot}}(t) - \rho_c(t)) \Rightarrow \\ &\frac{k}{a^2} = H^2(t)(\Omega(t) - 1). \end{aligned} \tag{2.36}$$

Finally, we write

$$\frac{1}{a(t)H(t)} = \sqrt{\frac{\Omega(t) - 1}{k}} \tag{2.37}$$

We can then calculate t_0 by writing dt as a function of H, as

$$t_0 = \int_0^{t_0} dt = \int_0^{a(t_0)} \frac{da'}{\dot{a}'} = \int_0^{a(t_0)} \frac{da'}{a'H'}. \tag{2.38}$$

2.3 Cosmological Evolution Formulas

Formulas as a Function of Redshift

We now define quantities as a function of the redshift. Since wavelengths in the expanding Universe are comoving (expand with the Universe),

$$\frac{\lambda(t)}{\lambda_0} = \frac{a(t)}{a_0}\,, \tag{2.39}$$

and then the redshift, which as found to be $1 + z = a_0/a(t)$, we obtain (remembering that $a(t)$ and $\lambda(t)$ correspond to the emission time for the wave)

$$1 + z = \frac{\lambda_{\text{detected}}}{\lambda_{\text{emitted}}} \equiv \frac{\lambda_0}{\lambda(t)} = \frac{a_0}{a(t)}. \tag{2.40}$$

Since frequencies satisfy $\nu = c/\lambda \propto 1/a$, we have

$$\nu a = \text{constant}. \tag{2.41}$$

The temperature in the Universe (which is in thermal equilibrium, at least as far as radiation goes) is defined by the black body radiation of the CMBR (cosmic microwave background radiation), which is a function of ν/T. The shape of the (Planck) distribution is obviously preserved as the Universe expands, which means that we also have

$$Ta = \text{const.} \Rightarrow T \propto \frac{1}{a(t)}. \tag{2.42}$$

We next express the age of the Universe as a function of z. Since $a_0/a(t) = 1 + z$, we replace in the formula for t (2.38)

$$x = \frac{a(t)}{a_0} = \frac{1}{1+z}. \tag{2.43}$$

We obtain

$$t = \int_0^{a_0/(1+z)} \frac{da'}{a'H'} = \frac{1}{H_0} \int_0^{1/(1+z)} \frac{dx}{x(H/H_0)}\,, \tag{2.44}$$

and we have to replace H/H_0 as a function of z, from $H = \dot{a}/a$,

$$a(t) = \frac{a_0}{1+z} = a_0 (t/t_0)^{\frac{2}{3(1+w)}}. \tag{2.45}$$

Specifically, the Friedmann equation is rewritten as (putting the curvature k/a^2, giving Ω_R on the same side as H^2)

$$H^2 = H_0^2 \left[\Omega_{0,m} \frac{\rho}{\rho_0} + \Omega_{0,R} \frac{a_0^2}{a^2} + \Omega_{0,\Lambda} \right]. \tag{2.46}$$

Using

$$\begin{aligned} \frac{\rho}{\rho_0} &= (a/a_0)^{-3(1+w)} = (1+z)^{3(1+w)} \\ \frac{a_0^2}{a^2} &= (1+z)^2\,, \end{aligned} \tag{2.47}$$

in H^2, and replacing H/H_0 in the formula for the age t, we obtain finally for the age of the Universe

$$t_0 = \frac{1}{H_0} \int_0^{1/(1+z)} \frac{dx}{\sqrt{\Omega_{0,\Lambda} x^2 + \Omega_{0,R} + \Omega_{0,m} x^{-(1+3w)}}}. \tag{2.48}$$

In particular, we saw in the previous chapter that for $\Omega_{\rm tot} = 1$, we have

$$t_0 = \frac{1}{H_0}\frac{2}{3(1+w)}. \tag{2.49}$$

Formulas as a Function of Conformal Time

We define the *conformal time* as the time when the metric is a time independent factor times an overall time dependent factor. Specifically,

$$d\eta = \frac{cdt}{a(t)}\ , \tag{2.50}$$

which means that

$$ds^2 = a^2(t)\left[-d\eta^2 + \frac{dl^2}{a^2(t)}\right]. \tag{2.51}$$

For instance, in FLRW coordinates, we have

$$ds^2 = a^2(t)\left[-d\eta^2 + \frac{dr'^2}{1-kr'^2} + r'^2(d\theta^2 + \sin^2\theta d\phi^2)\right]. \tag{2.52}$$

In terms of η, we obtain the table:

P	k	$t(\eta)$	$a(\eta)$	$H(\eta)$	$\Omega(\eta)$
$P = 0$ (M.D.)	1	$c^{-1}A(\eta - \sin\eta)$	$A(1-\cos\eta)$	$A^{-1}c\sin\eta(1-\cos\eta)^{-2}$	$2(1+\cos\eta)^{-1}$
	0	$c^{-1}A\eta^3/6$	$A\eta^2/2$	$A^{-1}4c\eta^{-3} = \frac{2}{3t}$	1
	-1	$c^{-1}A(\sinh\eta - \eta)$	$A(\cosh\eta - 1)$	$A^{-1}c\sinh\eta(\cosh\eta-1)^{-2}$	$2(1+\cosh\eta)^{-1}$
$P = \rho/3$ (R.D.)	1	$c^{-1}A'(\eta - \cos\eta)$	$A'\sin\eta$	$A'^{-1}c\cos\eta(\sin\eta)^{-2}$	$(\cos\eta)^{-1}$
	0	$c^{-1}A'\eta^2/2$	$A'\eta$	$A'^{-1}4c\eta^{-2} = \frac{1}{2t}$	1
	-1	$c^{-1}A'(\cosh\eta - 1)$	$A'\sinh\eta$	$A'^{-1}c\cosh\eta(\sinh\eta)^{-2}$	$(\cos\eta)^{-1}$

We will verify the cases with $\Omega = 1$ ($k = 0$), in which case as we saw

$$a(t) = a_0 = a_0(t/t_0)^{\frac{2}{3(1+w)}}\ , \tag{2.53}$$

so that

$$\eta(t) = c\int\frac{dt}{a(t)} \propto t^{\frac{1+3w}{3(1+w)}}. \tag{2.54}$$

Then for $w = 0$ (M.D.) we find

$$\eta(t) \propto t^{1/3} \Rightarrow t \propto \eta^3\ , \tag{2.55}$$

and for $w = 1/3$ (R.D.), we find

$$\eta(t) \propto t^{1/2} \Rightarrow t \propto \eta^2. \tag{2.56}$$

Moreover, since now

$$a(t) \propto t^{\frac{2}{3(1+w)}} \propto \eta^{\frac{2}{1+3w}} , \tag{2.57}$$

and for $\Omega = 1$ we have (2.49), which means also

$$H_0 = \frac{2}{3(1+w)t_0} , \tag{2.58}$$

for the $w = 0$ case (M.D.) we obtain

$$a(\eta) \propto \eta^2 \Rightarrow H_0 = \frac{2}{3t_0} , \tag{2.59}$$

whereas for the $w = 1/3$ case (R.D.), we obtain

$$a(\eta) \propto \eta \Rightarrow H_0 = \frac{1}{2t_0}. \tag{2.60}$$

We leave it as an exercise to check the cases for $k = +1$ and $k = -1$.

Important Concepts to Remember

- The expanding Universe is described by the FLRW metric, $ds^2 = -c^2dt^2 + dl^2$, with $dl^2 \propto a^2(t)$. In one expression, $dl^2 = a^2(t)d\vec{r}^2/[1 + kr^2/4]^2$. Here $k = 0$ for flat Universe, $k = +1$ for closed Universe and $k = -1$ for open Universe, compared to arbitrary k in the Newtonian approximation.
- A perfect fluid has $T^{\mu}{}_{\nu} = \text{diag}(-\rho, P, P, P)$, and a cosmological constant has $\rho_\Lambda = \Lambda/(8\pi G_N)$, and $P_\Lambda = -\rho_\Lambda$.
- The Friedmann equation is the same as in the Newtonian case, just that with the k as above, and the continuity equation is the same.
- Dust (pressureless) matter has $w = 0$, radiation has $w = 1/3$, stiff matter has $w = 1$ (for instance a scalar with $\phi = \phi(t)$ only), and cosmological constant has $w = -1$.
- The expansion age of the Universe is $t_0 = \int dt = \int da/aH = H_0^{-1}2/(3(1 + w))$.
- The temperature scales inversely with the scale factor, $T \propto 1/a$, as does the redshift, $1/(1 + z) = a(t)/a_0$.
- Conformal time is $d\eta = cdt/a(t)$, so that $a^2(t)$ is an overall (conformal) factor in the metric.

Further reading: Again most of the basic cosmology can be found in [1], see also the books [2, 3].

Exercises

(1) Calculate $t(\eta)$, $a(\eta)$ and $H(t)$ for $k = +1$ and $k = -1$, in the cases $w = 0$ and $w = 1/3$.
(2) Check that the Einstein equations on the FRLW metric reduce to the Friedmann equations (2.14) and (2.13).
(3) Assuming $\Omega \neq 1$ now, and a Universe made up of only matter and radiation, calculate $\Omega(t)^{-1}$ at $t \sim 10^{-10}s$. This is the "flatness problem".
(4) Show that the energy-momentum conservation $D_\mu T^\mu{}_\nu = 0$ is implied by the Einstein's equations, and that it implies the continuity (energy conservation) equation (2.24).

Chapter 3
The Propagation of Light and Measurements of Distance, Luminosity and Mass

In this chapter I will explain in some detail how measurements are made in astrophysics and cosmology, so that we can gain some faith in the fact that we actually can determine things precisely. I will start describing the propagation of light in a cosmological setting, follow on to say how we can use these formulas for distance measurements, and then optical measurements of distance, based on brightness and luminosity. These are based on the ideas of standard candles, so to understand that we will discuss shortly stellar evolution, and the possible stellar objects we can have. We will then classify the various measurements of distance, followed by possible determinations of mass of stellar objects.

3.1 Propagation of Light

The physical distance travelled by light, *at time* $t = t_2$ is (using the fact that light travels on $ds^2 = -c^2dt^2 + a^2(t)dr^2 = 0$)

$$\Delta l_{t=t_2} = a(t_2)\Delta r = a(t_2)c\int_{t_1}^{t_2}\frac{dt}{a(t)} = a(t_2)\Delta\eta. \tag{3.1}$$

We see that the distance is measured now, i.e., in terms of $a(t_2)$, and is proportional to the interval of conformal time.

In particular, we are interested in the case that $t_1 = 0$, i.e., in light that has travelled to us from the Big Bang. We apply to $k = 0$ (flat Universe), since this is the relevant case, and $P = 0$ (M.D.), since for most of its history, the Universe was matter dominated. Then we obtain

$$\Delta l_h = a(t_2)ct_2\int_0^{t_2}\frac{dt/t_2}{a(t_2)(t/t_2)^{2/3}} = 3ct_2. \tag{3.2}$$

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_3

This is then the *horizon distance*, i.e., the maximal region of the Universe that is knowable in principle (since nothing travels faster than the speed of light, and the earliest it can start is during the Big Bang).

Note that we are talking here about the so-called *particle horizon*, which means what we can see from particles that travelled to us since the Big Bang. But there are other notions of horizon. One of them is the *(future) event horizon*, which is the largest (comoving) distance from which a particle sent now will even (in the future) reach us. Thus, in the case that the Universe goes on forever (and does not end, or collapse), this is the complement of the particle horizon,

$$R_{\text{ev.hor.}} = a(t)c \int_t^\infty \frac{dt'}{a(t')} \tag{3.3}$$

If the Universe ends at t_e, we only integrate up to t_e. There is also the notion of *Hubble horizon*, which is what we obtain in the linear Hubble law if we put the maximum velocity possible, $v = c$, so

$$d_H = \frac{c}{H}. \tag{3.4}$$

For a power law expansion, $a(t) \propto t^p$, all the horizons are of the same order, differing only by number factors, so we will not always be too careful of the distinctions between the horizons from now on.

In the case $k = 0$, but $P = \rho/3$ (R.D.), which was valid for a short while before the M.D. era, we have the (particle) horizon distance

$$\Delta l_h = ct_2 \int_0^{t_2} \frac{dt/t_2}{(t/t_2)^{1/2}} = 2ct_2. \tag{3.5}$$

This distance defines then, for all practical purposes, the size of the Universe (the Universe may be larger than that, but we wouldn't know, since this is the knowable Universe).

Note the strange fact that $l_h > ct_2$, so it seems like "the speed of light is greater than the speed of light". But this is of course due to the fact that l_h is measured *today*, and the Universe expands as light travels.

Another observation is that there is no global inertial reference frame, so we cannot use special relativity. That is as it should be, since we are not in Minkowski space, but in the FLRW Universe. In this space, we have a preferred time t that we have been using, defined by the cosmological expansion, and it defines a preferred reference frame (which is however not inertial).

Because our observations will have to do with light propagating from some distance place, at its constant speed, until now, we can use as a measurement of distance the redshift z, since z corresponds to a time, when we have a scale $a(t)$, from which light has propagated to us. More precisely, we have $1 + z = a_0/a(t)$. As we saw in the previous chapters, z was defined in terms of the frequency shift as

$$z = \frac{\nu_{\rm em} - \nu_{\rm det}}{\nu_{\rm det}} \ , \tag{3.6}$$

and $\nu \propto 1/a$, as well as the temperature, $T \propto 1/a$, leading to

$$T \propto 1 + z \propto a \Rightarrow \frac{T}{T_0} = 1 + z = \frac{\nu}{\nu_0} \equiv \frac{\nu_{\rm em}}{\nu_{\rm obs}}. \tag{3.7}$$

3.2 Measurements of Distance by Light

In Euclidean space, the angular size θ is related to the linear size l and the distance to the object r by

$$\theta = \frac{l}{r}. \tag{3.8}$$

In non-Euclidean space however, it is not so simple. But reversely, we can *define* a distance called *angular size distance*, which is implicitly a function of the redshift z at which the source is located, by

$$\bar{r} \equiv \frac{l}{\theta} = f(z). \tag{3.9}$$

Using the coordinate system where we had the FLRW metric

$$ds^2 = -c^2 dt^2 + a^2(t)[dr^2 + g_k(r)^2(d\theta^2 + \sin^2\theta d\phi^2)] \ , \tag{3.10}$$

where $g_k(r) = (r, \sinh r, \sin r)$ for $k = 0, -1, +1$, we have seen that a circle in the sky, defined by θ, has the circumference

$$S_k = 2\pi g_k(r) a(t). \tag{3.11}$$

For $\Delta\theta$ replacing 2π, S_k corresponds to the linear size $l = \Delta\theta\bar{r}$, so we have

$$\bar{r} = a_0 g_k(r). \tag{3.12}$$

As for r, it can be determined as before, from the fact that light propagates radially towards us, on $ds^2 = -c^2 dt^2 + a^2(t) dr^2 = 0$, so

$$r = c\int \frac{dt}{a} = c\int \frac{da}{a\dot{a}} = \frac{c}{H_0}\int \frac{da}{a^2 H/H_0(z)} \ , \tag{3.13}$$

is implicitly a function of the redshift z. Finally, we define the function $\psi(z)$ by

$$\bar{r} \equiv \frac{c}{H_0}\psi(z). \tag{3.14}$$

For pure dust matter ($P = 0$, M.D.), we obtain

$$\psi(z) = \frac{2}{\Omega_0^2(1+z)^2}\left\{\Omega_0 z + (\Omega_0 - 2)[(1+\Omega_0 z)^{1/2} - 1]\right\} , \tag{3.15}$$

which is left as an exercise to prove.

As a function of the variable

$$\Delta \equiv \frac{z}{1+z} , \tag{3.16}$$

which takes values from 0 to 1 (for $z = \infty$, at $t = 0$), and at fixed linear size l, the angular size θ *for* $\Lambda = 0, \Omega > 1$ (not the case with the current cosmology, of course) has a *minimum*. That is, as Δ increases, θ goes to a minimum, and then increases again close to 1, all the way to infinity!

Of course, we cannot see arbitrarily far in the past, i.e., arbitrarily close to the Big Bang (for $z \to \infty$). In particular, there is a maximum z, corresponding to the "surface of last scattering", when photons decoupled from matter, and atoms formed. Before that, the Universe is opaque (after that, it is transparent), and moreover there are no structures to be seen, not even atoms. So because of that, there are no large angle objects to be seen, but otherwise we would see large angles for things very close to the Big Bang.

3.3 Optical Measurements

We move on to optical measurements and their use for distance measurements.

In flat space, if there is no absorption, nor scattering, the brightness of a source is independent on the distance and on the existence of refraction.

In curved space, the analogous statement is different: the brightness of a source depends on z (and so, on the distance), but not on the curvature, nor on the density ρ.

The *brightness* is defined as the power P (energy per unit time), per unit solid angle and perpendicular area,

$$I = \frac{d^2 P_{\rm obs}}{d\Omega dS_\perp}. \tag{3.17}$$

Note that here $d\Omega$ means $d\Omega_{\rm em}$, which is the solid angle of the source viewed from the observer, whereas $dS_\perp$ means $dS_{\rm obs}\cos\theta$, i.e., the area of the observer, perpendicular to the direction of propagation. But it doesn't matter, since we have the equality

$$B = \frac{d^2 P_{\rm obs}}{d\Omega_{\rm obs} dS_{\rm em}\cos\theta} = I = \frac{d^2 P_{\rm obs}}{d\Omega_{\rm em} dS_{\rm obs}\cos\theta}. \tag{3.18}$$

We can also define the *spectral brightness*, as the brightness per unit of frequency,

$$F(\omega) = \frac{d^3 P_{\text{obs}}}{d\Omega dS_\perp d\omega} \ , \tag{3.19}$$

and then for blackbody radiation of radiation temperature T, the spectral brightness $F(\omega, T)$ is invariant. Moreover, as we already observed, the temperature decays with the expansion of the Universe, so

$$T_{\text{obs}} = \frac{T_{\text{em}}}{1+z}. \tag{3.20}$$

Then the brightness satisfies a law similar to Stefan–Boltzmann,

$$I = \text{const.} \times T^4 = \frac{I_0}{(1+z)^4}. \tag{3.21}$$

The *luminosity* L is defined as the total energy emitted by the source per unit time, in the rest frame. Since $\theta = l/\bar{r} = l/(c\psi(z)/H_0)$, for light propagation, the solid angle $d\Omega$ of the detector viewed from the source is, in terms of the area dS of the detector,

$$d\Omega = \frac{dS}{(c\psi(z)/H_0)^2} \ , \tag{3.22}$$

which means that the total area at the distance of the detector is

$$S = 4\pi\bar{r}^2 = 4\pi(c\psi(z)/H_0)^2. \tag{3.23}$$

Define the *total measured (bolometric) flux*, i.e., energy per unit time per unit surface of the detector. Then from the above it is given in terms of the luminosity of the source at emission L_{em} by

$$\mathcal{F}_{\text{obs}} = \frac{L_{\text{em}}}{4\pi\bar{r}^2(1+z)^4} = \frac{L_{\text{em}}}{4\pi(c\psi(z)/H_0)^2(1+z)^4} \ , \tag{3.24}$$

where the factor $(1+z)^4$ is the same factor as in the brightness I, and comes from the variation of the luminosity from emission to detection.

We can define the *luminosity distance* d_L as the distance in Euclidean space producing the same $\mathcal{F}_{\text{obs}}$ at the same $\mathcal{L}_{\text{em}}$, i.e.,

$$\mathcal{F}_{\text{obs}} = \frac{L_{\text{em}}}{4\pi d_L^2} \Rightarrow d_L = \frac{c}{H_0}\psi(z)(1+z)^2. \tag{3.25}$$

Note that in terms of the luminosity *today*, $L = L_{\rm em}/(1+z)^2$, we have

$$\mathcal{F}_{\rm obs} = \frac{L}{4\pi \bar{r}^2 (1+z)^2}. \tag{3.26}$$

We can also define something that is directly observed, basically $\mathcal{F}_{\rm obs}$ on a strange scale, called the *apparent magnitude* of an object,

$$\begin{aligned} m &= -2.5 \log_{10} \mathcal{F}_{\rm obs} + \text{const.} \\ &= 5 \log_{10} d_L + \text{const.}' \end{aligned} \tag{3.27}$$

where in the second equality we have expressed $\mathcal{F}_{\rm obs}$ in terms of d_L, so that now $L_{\rm em}$ is absorbed in the constant.

Finally, define the *absolute magnitude* M, as the apparent magnitude m of an object situated at 10 parsecs (about 3×10^{19} cm) away. Therefore, since d_L is now fixed, the absolute magnitude M is a measure of the luminosity $L_{\rm em}$.

Substituting the form of $\psi(z)$ for the M.D. case into the formula for d_L (3.25), we find

$$H_0 d_L = \frac{2}{\Omega_0^2}\{\Omega_0 z + (\Omega_0 - 2)[(1+\Omega_0 z)^{1/2} - 1]\}. \tag{3.28}$$

Then the apparent (bolometric) magnitude is given by

$$\begin{aligned} m &= 5 \log_{10}(H_0 d_L) + C_1 \\ C_1 &= M_{\rm bol,em.} - 45.06 - 5 \log_{10} H_0. \end{aligned} \tag{3.29}$$

We intend to calculate this relation for nearby objects, for which we can expand $a(t)$ to second order, using H_0 and q_0 as the coefficients of the Taylor expansion, so

$$\frac{1}{1+z} = \frac{a(t)}{a(t_0)} = 1 + H_0(t - t_0) - \frac{1}{2} q_0 H_0^2 (t - t_0)^2 + \cdots \tag{3.30}$$

From it, we derive

$$z \simeq -H_0(t - t_0) + \left(1 + \frac{q_0}{2}\right) H_0^2 (t - t_0)^2 + \cdots, \tag{3.31}$$

which inverts as

$$t_0 - t \simeq H_0^{-1} \left[z - \left(1 + \frac{q_0}{2}\right) z^2 + \cdots \right]. \tag{3.32}$$

On the other hand, expanding the relation $r = c \int dt/a(t)$ to second order in the $k = 0$ (flat) case, we find

$$\bar{r} = a^{-1}(t_0) \left[(t_0 - t_1) + \frac{H_0}{2}(t_0 - t_1)^2 + \cdots \right] \tag{3.33}$$

and replacing in it the $t - t_0$ as a function of z, we obtain

$$\bar{r} = a^{-1}(t_0) H_0^{-1} \left[z - \frac{1}{2}(1 + q_0) z^2 + \cdots \right]. \tag{3.34}$$

Finally replacing this in $d_L = \bar{r}(1 + z)^2$, we find

$$H_0 d_L = z \left(1 + \frac{1}{2}(1 - q_0) z + \cdots \right) , \tag{3.35}$$

and the bolometric apparent magnitude is

$$\begin{aligned} m &\simeq 5 \log_{10} z + 1.086(1 - q_0) z + C_1 \\ C_1 &= M_{bol} + 45.06 - 5 \log_{10} H_0. \end{aligned} \tag{3.36}$$

3.4 Stellar Evolution

At this moment, we describe shortly stellar evolution, since it is important in order to understand the measurements of distance and mass. About 100 years ago (about 1910), a diagram was devised that encapsulates very well stellar evolution, called the *Hertzsprung–Russel* diagram, after its discoverers, see Fig. 3.1. It maps luminosity (or absolute magnitude M, which is a measurement of it, as we saw), usually rescaled, so $L_{\text{em}}/L_{\text{Sun}}$, versus the temperature T, or rather, the colour. Note that the temperature is represented in reverse in the diagram (increasing from right to left), or colour from Blue to Red. The diagram is found from nearby stars, for which we can measure M, and we measure T from the blackbody radiation, but we assume it is valid for all stars.

We then observe a startling regularity: most stars (about 90%) fall onto a descending line in the Herzsprung–Russel diagram, called the *main sequence*. Above the main sequence, we have some lines representing red giants, below some lines for white dwarfs. It is found that the luminosity also increases with the mass, and we can represent the mass of stars on the diagram. Then the main sequence has about $M = 0.1 M_{\text{Sun}}$ at the lower end, and about $M = 10 M_{\text{Sun}}$ at the upper end, with some stars going until about $M = 60 M_{\text{Sun}}$.

Red giants and white dwarfs are just later stages of stellar evolution, as we will see shortly. We have also variable stars, in particular *Cepheid variables* after the first discovered, Delta Cephei. These are stars that have a period of oscillation of days or months, for up to 30% of intensity. They are stars in later states of evolution as well, in which they collapse or expand repeatedly. The most important example is the North Star (Polaris), that oscillates up to 9% every 4 days. Cepheid variables are localized in a region above the main sequence also. Finally, we also have *novas*, which are stars that flare up every 30–50 years, by a factor of 10^6 times, over a

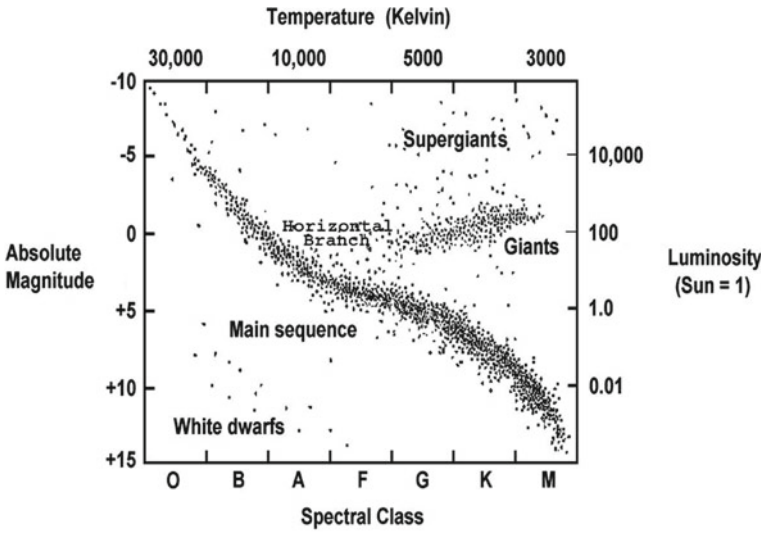

Fig. 3.1 Herzsprung–Russel diagram (Image: NASA/Chandra)

short time. They are stars in binaries, together with a white dwarf, and the accretion produces instabilities.

The evolution of stars can be described depending on their mass:

- If $M < 4M_{\text{Sun}}$, when its fuel is consumed, the star becomes a *red giant* and explodes in a "Helium flare-up", i.e., an explosion fueled by (nuclear) reacting Helium. The explosion leaves a corona (planetary nebula), and a core, which becomes a white dwarf, with a mass $< 1.4M_{\text{Sun}}$.
- If $4M_{\text{Sun}} < M < 8M_{\text{Sun}}$, the star becomes a white dwarf, with a mass $> 1.4M_{\text{Sun}}$. Then the electron degeneracy pressure (incompressibility of electrons) eventually cannot keep the star from collapsing, and it has a "Carbon flare-up", i.e., an explosion fueled by (nuclear) reacting Carbon. The explosion is a *supernova* explosion, specifically called a type I Supernova.
- If $M > 8M_{\text{Sun}}$, there is also a supernova explosion, but it has no Carbon flare-up. Instead, it produces heavier elements, and then the star collapses under gravity until the nuclei touch each other, and then explodes, creating even more heavy elements. After that, the core becomes a *neutron star*.
- A heavy enough neutron star will collapse gravitationally to a black hole (which afterwards can accrete anything nearby).

3.5 Measurements of Distance

We now turn to the various measurements of distance that we can use. The reason we can measure distances all the way to cosmological distances (sizes of the order of the Universe's) is that we have various methods that are valid *on different, yet*

overlapping, distance intervals. By verifying or calibrating the method on the overlap, we can then use it at larger distances, etc.

(a) **Trigonometric Parallax**

At the smallest scales, we use a direct measure of distance, based on the fact that the Earth moves along the year around the Sun (on the ecliptic), and in doing so, the angle to a nearby star changes slightly by a $\delta\alpha$. This manifest itself in a change $\delta\alpha$ of the position on the stars on the sky (relative to very distant objects, "at infinity") as one moves on diametrically opposite positions on the ecliptic, see Fig. 3.2.

This method gives the definition of a *parsec*, as the distance at which the parallax is of one arc second. It equals about 3.26 light years. The method is used until about 100 light years.

(b) **Moving Cluster Method**

Consider a cluster of stars. Then individual stars in the cluster move at velocity v at an angle α from the direction to the star. We can then measure the radial component of the velocity, $v_r = v\cos\alpha$, through the Doppler shift. We can measure the average angle α of the stars in the cluster as follows: we can measure an angular distance on the sky of the various stars, and the angular velocities, extrapolated, (statistically) converge on a point in the sky situated at an angular distance α from the cluster, see Fig. 3.3. Then we calculate the tangent velocity $v_T = v_r \tan\alpha$. Measuring also the angular velocity on the sky μ ("proper motion"), measured over a long period, we can find the distance as

$$d = \frac{v_T}{\mu}. \tag{3.37}$$

This method is used up to about 500 light years. In particular, it is used for the Hyades cluster, situated at 120 light years, and the Scorpio–Centaurus association, situated at about 500 light years.

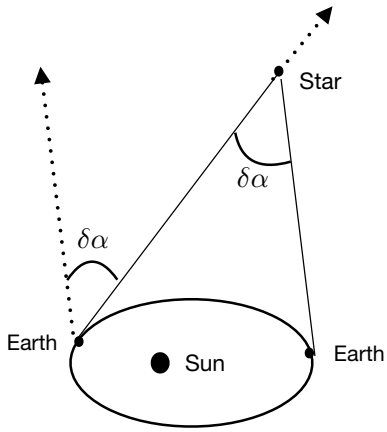

Fig. 3.2 Trigonometric parallax: the motion of the Earth around the Sun generates a corresponding motion of nearby stars on the sky

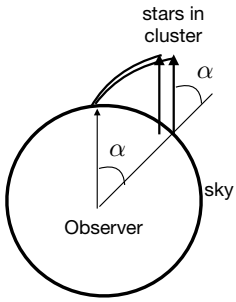

Fig. 3.3 Moving cluster method: the tangential (on the sky) velocities of the moving cluster (statistically) converge on a point in the sky, corresponding to the angle α of the velocity vector with the radial direction

(c) **Spectroscopic Parallax**

It is used for stars on the *main sequence*, for which there is collisional broadening of the spectral lines. Then one measures the apparent magnitude m, and the temperature T from its spectrum. Then, derive the absolute magnitude M from the Hertzsprung–Russel diagram. This allows us to calculate the distance from the relation between m and M,

$$d(\text{parsecs}) = 10^{\frac{m-M+5}{5}}. \tag{3.38}$$

This method is of course not very accurate, since it relies on the fact that stars lie on the Main sequence, which is not very precise (the main sequence is experimentally a quite widened line). But the advantage is that the accuracy remains constant, of about 15%, independent of the distance. This allows it to be used for distances larger than 100 light years.

(d) **Main Sequence Fitting**

It is a variant of the above. One applies the method above to *clusters of stars*, and then calibrates first by another method for nearby clusters.

(e) **Globular Clusters**

It is used for extragalactic distances (distances outside the Milky Way). One considers clusters of stars (of population II stars, on a line parallel to the Main Sequence line). We calibrate it for nearby population II stars, then use it for globular clusters in the Milky Way, then extend it to other galaxies. It is used up to about 30 million light years.

(f) **RR Lyrae Stars, or the Statistical Parallax Method**

The method is used for RR Lyrae variable stars, which are found to be *standard candles*. One measures the proper motion of the star and the peculiar motion (extra, excluding the average motion of the stars). As before, the radial component is measured by Doppler shift, and averaging over the velocities of several stars, one finds $\langle v_r \rangle = \langle v_\theta \rangle = \langle v_\phi \rangle$. Then proceed as in the moving cluster method, and find that the absolute magnitude M of RR Lyrae is always close to 0.6, i.e., the RR Lyrae are standard candles, and can be used in galaxies where they are visible.

(g) **Cepheid Variables**

The Cepheid variables are brighter than the RR Lyrae. In a similar manner, one finds that the visual absolute magnitude M is linear with the period P of the variability. That in effect acts as another type of standard candles, determining M from the easily measurable period. The method is used up to about 10 million light years.

(h) **Bright Blue Stars**

It is assumed that the brightest blue stars are blue supergiants, and have the same luminosity in all galaxies (there is a cut-off in luminosity there), $L \sim 10^6 L_{\rm Sun}$.

(i) **Size of H II Regions**

The sphere of ionized matter around bright blue stars is found to have a radius of about 150 light years. These are the H II regions, correlated to the bright blue star at the center, and have a standard width, allowing us to calculate the distance to them.

(j) **Supernovae**

It is found that the peak luminosity of a supernova explosion is about the same as the one of a bright galaxy, in all galaxies. That allows us to treat them at standard candles. The supernovae close by are spectacular: the most famous, in 1054, occurred in our own Milky Way galaxy, and was recorded in China, as being bright even during daytime. It left behind the Crab Nebula, of (still moving at relativistic speeds) ejecta from the explosion.

(k) **Galaxies**

It is found that the brightest member of a cluster of galaxies has the same luminosity L, i.e., there is a cut-off, so it can be used as a standard candle. It is a method used up to about 2 billion light years.

(l) **Redshift**

Finally, at the largest distances, we can only used the redshift for a measure of distance, assuming a cosmological model for the expansion.

3.6 Mass Determination for Distant Bodies

We finally describe methods of measurement of mass.
(a) There is an empirical law for M versus $L_{\rm em}$, from models.
(b) There is also a theoretical M versus $L_{\rm em}$ relation, from models.
(c) **Stellar spectra.**

We can measure the surface gravity of a star from the widths of spectral lines (which normally are associated with the decay of the state). The gravity at the surface of the star, where the light is emitted, influences the widths. Then the surface gravity is

$$g = \frac{MG_N}{R^2} \ , \tag{3.39}$$

and the luminosity is given by the Stefan–Boltzmann law,

$$L = 4\pi R^2 \sigma T^4 \ , \tag{3.40}$$

resulting in the relation

$$\frac{M}{L} = \frac{g}{4\pi G_N \sigma T^4}. \tag{3.41}$$

Measuring T and L (and knowing g), we can find M.

(d) **Pulsating Stars**

For the pulsating star, the period P is related to the average density $\bar{\rho}$ by the relation

$$P = 0.06\sqrt{\frac{\bar{\rho}_{\rm Sun}}{\bar{\rho}}} \tag{3.42}$$

from which we determine $\bar{\rho}$. Determining L (for the Cepheids from the $P - L$ relation, for the others from distance), and $L = (4\pi R^2)\sigma T^4$, leading to a measurement of R, after which we have simply

$$M = \frac{4\pi}{3}\bar{\rho}R^3. \tag{3.43}$$

(e) **Gravitational Redshift**

The redshift due to the gravity at the surface of a dense object is

$$z_{GRS} \simeq \frac{2MG_N}{Rc^2}. \tag{3.44}$$

In white dwarfs, z_{GRS} can be up to $10^{-5} - 10^{-4}$, and if we measure the redshift and R, we can measure M.

Important Concepts to Remember

- The physical distance travelled by light at t_2 is $\Delta l = a(t_2)\Delta\eta = a(t_2)c\int dt/a(t) > ct_2$.
- The angular size distance is $\bar{r} = l/\theta = f(z)$, where l is linear size and θ is angular size.
- For light propagation, we have $\bar{r} \equiv c\psi(z)/H_0$.
- The luminosity distance d_L is the distance in Euclidean space producing the same observed flux $\mathcal{F}_{\rm obs}$ for a given luminosity $L_{\rm em}$ at emission, so $\mathcal{F}_{\rm obs} = L_{\rm em}/(4\pi d_L^2)$, so $d_L = (c/H_0)\psi(z)(1+z)^2$.
- The apparent magnitude is $m = -2.5\log_{10}\mathcal{F}_{\rm obs} + \text{const.} = 5\log_{10} d_L + \text{const.}$.

- The absolute magnitude M is the apparent magnitude of an object 10 parsecs away, and thus measures L_{em}.
- To linear order $H_0 d_L = z\left(1 + \frac{1}{2}(1 - q_0)z + \cdots\right)$.
- Most stars are on the main sequence, for luminosity (or absolute magnitude) versus temperature (or colour).
- For $M < 4M_{\text{Sun}}$, the star becomes a red giant, and explodes in a He flare-up leaving a white dwarf. For $4M_{\text{Sun}} < M < 8M_{\text{Sun}}$, the resulting white dwarf has a Carbon flare-up giving a supernova explosion. If $M > 8M_{\text{Sun}}$, there is no Carbon flare-up, but the supernova produces heavier elements, and it creates a neutron star. A large enough neutron star collapses to a black hole.
- Measurements of distance are progressive, from small distances (where we measure them almost directly, like in the parallax method), to large distances, using overlapping domains of validity.
- There are a variety of mass measurements as well.

Further reading: Some of the things I said can be found in [1, 4], others in astrophysics textbooks.

Exercises

(1) Prove that in the M.D. case, the angular size distance is

$$\frac{c}{H_0}\psi(z) = \frac{c}{H_0}\frac{2}{\Omega_0^2(1+z)^2}\{\Omega_0 z + (\Omega_0 - 2)[(1+\Omega_0 z)^{1/2} - 1]\}. \tag{3.45}$$

(2) Write a formula for the future event horizon R_h, the maximal physical distance a fixed observer may come to observe. Calculate it approximately for the current, Λ dominated period.
(3) Calculate the particle horizon for an Universe with $\Omega = 1$, currently containing 30% matter and 70% cosmological constant, as a function of h_0, the Hubble constant today.
(4) Estimate the dimensionless gravitational potential at the surface of the Earth.

Chapter 4
Evidence for Dark Matter and the ΛCDM Model

In this chapter we will present the evidence for Dark Matter, and the resulting "Standard Model of Cosmology," the ΛCDM model (cosmological constant plus cold dark matter). Dark Matter is what the name suggests, matter that is dark, i.e., non-luminous. That is, matter that we cannot detect from emitted light. I will present the evidences, then review the possible candidates for dark matter, and finish describing the ΛCDM model.

4.1 Evidences for Dark Matter

I will present the various evidences as a list, starting with the first ones to be observed.

(1) **The Coma Cluster of Galaxies** (situated at about 20 Mpc distance)

The first evidence for dark matter dates back almost a century, to 1933, when *Zwicky*, studying the Coma Cluster, noticed an unusual thing. The Coma Cluster is made up of many galaxies, interacting gravitationally, that can be for all intents and purposes considered as point-like. This system then is no different than a gas in kinetic theory, with interactions generated by some potential.

In that case, the *virial theorem* states that for a gravitationally bound system that has virialized, i.e., "thermalized" (in the language of kinetic theory), on the statistical (ensemble) average, defined *locally*, we should have

$$2\langle T\rangle = -\langle U\rangle \Rightarrow \langle v^2\rangle \sim \frac{G_N M}{r}. \tag{4.1}$$

Here $v = v_p = v_{\text{tot}} - Hr$ is the peculiar velocity of the galaxies, i.e., the total velocity minus the Hubble expansion. But taking for M the *visible* mass, Zwicky found that

$$v^2 \gg \frac{G_N M_{\text{visible}}}{r}. \tag{4.2}$$

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_4

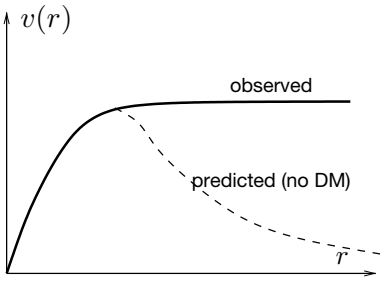

Fig. 4.1 Rotational curves of galaxies. Without the dark matter, outside the radius of the galaxy, the velocity drops as $v \sim 1/\sqrt{r}$ (dotted), but with dark matter (observed) it is approximately flat (full line)

That suggests that there should be extra matter that is dark (not visible), called *dark matter*, and that

$$M_{\text{dark}} \gg M_{\text{visible}}. \tag{4.3}$$

But this hypothesis was discarded untill the 1970s, assuming that it was a statistical fluke, since it was an isolated phenomenon,

(2) **The Rotational Curves of Galaxies**

In the 1970s, Ford and Rubin discovered that rotational curves of galaxies are flat. This is similar to the previous case, just that instead of clusters of galaxies, one considers a single galaxy, specifically a spiral galaxy like our own Milky Way, and stars orbiting around its center of mass. Besides the bulk of the stars, forming the characteristic spiral shape, there are outliers, that live outside the visible matter.

For such a star, orbiting the center of the galaxy and on *bound trajectories* (like the planets around the Sun, say), we have a relation that is now true independently of statistics (for each individual star, unlike the case of the cluster of galaxies), equaling the centrifugal and centripetal force,

$$F_{\text{centrifugal}} = \frac{mv^2}{R} = \frac{mG_N M(r)}{r^2} = F_{\text{grav}}\ , \tag{4.4}$$

from which we derive that

$$v = \sqrt{\frac{G_N M(r)}{r}}. \tag{4.5}$$

We look for stars situated at a distance greater than the location of most visible mass (outliers). In that case, we expect that $M(r) \simeq M$ =const., so $v \propto \frac{1}{\sqrt{r}}$, but instead one observes $v \simeq \text{const}(r)$, as in Fig. 4.1. That implies

$$M(r) \propto r! \tag{4.6}$$

Note that this is not what we would expect from a constant density of matter, which would give $M(r) \propto r^3$, but rather we have some diffuse matter, that is not

completely bound to the galaxy, forming a *halo*, yet is still centered on the center of the galaxy. This would be *dark matter*.

The best fit for the data is then (cold) dark matter, a type of matter that doesn't interact with normal matter, except gravitationally (since gravity interacts with all matter equally). It could have some self-interaction, though that is not clear. Due to the gravitational interaction, it has the same center of mass as the galaxy, in the case of rotation curves. The fact that is cold (very low temperature) will be explained better later.

As an example, consider the galaxy NGC 6503, where the matter from the disk gives a sharp peak in the centrifugal velocity v_c at about 3 Kpc (kiloparsecs) at close to 10 km/s, then decays sharply, until about 20 Kpc. The Dark Matter halo reaches a maximum of v_c at about 10 Kpc, after which if flattens out, so that the total v_c rises quickly until about 120 km/s at about 3 Kpc, after which it stays really flat.

The limitation in distance away from the center of the galaxy is how far away can we see the light from the star, or neutral H, giving out 21 cm waves, which is about tens of Kpc.

As we said, the velocity field v_c is really flat, but that is not reproduced really perfectly from Cold Dark Matter simulations. In fact, interestingly, an alternative theory called MOND (described in an exercise) gives a better fit, though that theory cannot explain well other evidences for Dark Matter (like the Bullet Cluster below).

(3) **Hot Gas in Clusters** (for instance, in the Coma Cluster)

The principle is similar to the case of the Coma Cluster of galaxies, except we track the mass through the hot gas moving around. Its velocity (squared) v^2 is measured from the temperature of the hot virialized gas, and as before we compare with GM_{visible}/r.

(4) **The Bullet Cluster**

These are two colliding clusters of galaxies, where one goes through the other, so it is perhaps a misnomer, but it is associated with the fact that there is a lot of hot plasma (gas) remnant of the passing of one cluster through the other, like a bullet passing through some material and leaving a hot trail behind it. The interesting thing is that most visible mass is situated in the plasma, which is in between the galaxies, since it was slowed down in the collision. The center of mass of the visible mass is then in between the galaxies, but the center of mass obtained from gravitational lensing that, as we will shortly see, maps out all matter, is different. Dark matter actually traces the galaxies, so it is centered on the two galaxies. In fact, one obtains

$$\frac{M_{gas}}{M_{DM}} \sim \frac{1}{6-7} \,, \tag{4.7}$$

as obtained in the ΛCDM model.

This is one of the most difficult tests to reproduce outside of the Dark Matter idea (for alternatives to Dark Matter), because of its qualitatively different nature. It is usually stated that MOND has a problem in reproducing it (though in its relativistic

version, TeVeS, it is sometimes claimed that the mismatch is reduced to a factor of 2). Similarly for modified gravity models.

(5) **Surveys of Visible Matter**

By taking large scale surveys of the visible matter, we can map out the distribution in the Universe. The first one was done by J. Oort in 1956, but the results have not changed much since then. One computes the ratio of mass to luminosity M/L for different galaxies, and finds that on the average

$$\left\langle \frac{M}{L} \right\rangle \sim 21 \frac{M_{\text{Sun}}}{L_{\text{Sun}}}. \tag{4.8}$$

Then one maps out the luminosities of galaxies, and from the above ratio we find the masses. Adding them up we find

$$\Omega_{\text{visible}} \sim 0.03. \tag{4.9}$$

(6) **Gravitational Lensing**

In general relativity, massive objects like stars and galaxies deflect light. Famously this is one of the first decisive tests of general relativity, when deflection of the light from the stars behind the Sun was measured during an eclipse, and found to agree with general relativity rather than special relativity (the difference between them is a factor of 2).

The angle of deflection of light (between the original and the final asymptotic lines) is

$$\delta\phi = \frac{4G_N M/c^2}{r}\ , \tag{4.10}$$

where r is the minimum distance at which light passes from the object as it is deflected.

The deflection means that various rays of light deflected by the object in front of them can get focused on the same observer, like in the case of a lens, hence the name *gravitational lensing*. Of course, unlike the case of a lens, the gravitational deflection is gradual (gravity being a long range force), it is not completely localized. From the point of view of the observer, he would see the light from the object behind the deflector as an arc or arcs centered on the deflector. The deflection angle observed on the sky (angular distance between the arc and the true position), i.e., the deflection angle from the point of view of the observer, is

$$\alpha = \sqrt{\frac{4G_N M}{d}}\ , \tag{4.11}$$

where d is the impact parameter (the minimal distance it would have passed from the deflector, were it not deflected), for small angles approximately equal to d', the smallest distance passed from the deflector, see Fig. 4.2.

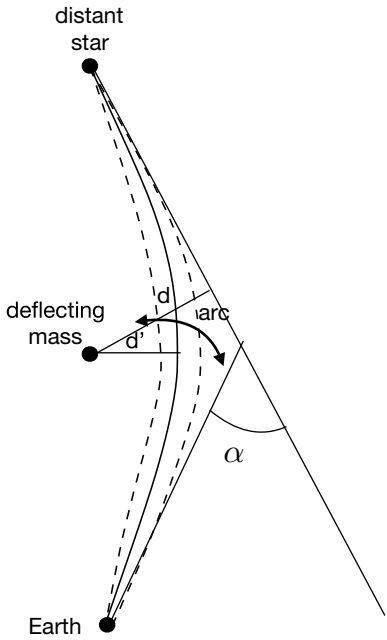

Fig. 4.2 Gravitational lensing: Light from a distant star is deflected by a deflecting mass (perhaps a local star), creating an arc on the sky, instead of a point. The angle α depends on the impact parameter d, or the smallest distance to the deflector d'

By applying this method to a cluster of galaxies (situated at tens of Mpcs), and looking at other clusters behind it (also tens of Mpcs), in the form of arcs in the field of vision, we can map out the mass distribution.

(6) **MACHOs** (MAssive Compact Halo Objects)

A kind of subcategory of gravitational lensing is MACHOs, since they are an alternative explanation for why there would be Dark Matter, and they are also related to gravitational lensing. These would be large, non-luminous objects that, as they pass in front of various sources that we see, give a very specific signal, namely *achromatic* (independent of frequency, since it is a gravitational effect) brightening. The brightening is due to the same phenomenon as the gravitational lensing, and is not possible for a non-gravitational source.

But these were not seen to the degree expected if they were supposed to in order to be an alternative to dark matter, hence they are excluded as an explanation. More on them later.

(7) **Peculiar Velocities of Galaxies in Clusters**

Considering a cluster of galaxies, one can measure the individual (peculiar) velocities

$$v_p = v_{\text{obs}} - Hr \tag{4.12}$$

of galaxies, and one makes a map of the velocity flow, reconstructing the gravitational potential. A formula that is found to work relates the divergence of the peculiar

velocity field to the overdensity of galaxies $\delta\rho/\rho$ by

$$\vec{\nabla} \cdot \vec{v} = -\Omega^{0.6} \frac{\delta\rho}{\rho}. \tag{4.13}$$

(8) **Big Bang Nucleosynthesis**

As we have already suggested, when we go back in time, the last moment that we can experimentally measure cosmology precisely is the Big Bang Nucleosynthesis, the period when nuclei get formed. After that, there is the moment of recombination, when atoms get formed, and the Universe becomes transparent to light, and the CMBR is formed.

During Big Bang Nucleosynthesis (BBN), the light elements are formed (the heavier ones are formed in nova and supernova explosions), mostly H, so it is used as reference. The abundances of He/H, $(D + {}^3He)/H$, Li/H created during BBN are matched against observed abundances in the Universe (which come mostly from BBN, and to a smaller degree from nuclear reactions within stars), so we have a precise test of the conditions at BBN. One of the factors that influence the abundances is Ω_{baryons}, and one finds a

$$\Omega_{\text{baryons}} \sim 0.05. \tag{4.14}$$

A more precise description of Big Bang Nucleosynthesis (BBN) will be given in Chap. 6.

(9) **CMBR**

Finally, of course, we have the CMBR. As we already said, it fixed Ω_{total} to be 1, which together with the measurement of acceleration from type I_a Supernovae means that $\Omega_{\text{matter}} \sim 0.3$, implying there is dark matter through the above arguments.

But the CMBR depends also on Ω_{baryons}, and one finds

$$\Omega_{\text{baryons}} \sim 0.05. \tag{4.15}$$

Actually, it is $\Omega_{\text{baryons}} h^2$ that is the parameter that influences it ($H = h \cdot 100\,\text{km/s/Mpc}$), but given other measurements of H, this is what we obtain.

In conclusion, we obtain

$$\Omega_{\text{total}} = 1; \quad \Omega_\Lambda = 0.73; \quad \Omega_{\text{DM}} = 0.23; \quad \Omega_{\text{baryons}} = 0.04. \tag{4.16}$$

4.2 Candidates for Dark Matter

Having established the existence of Dark Matter, we now consider various candidates for what they can be.

Cold Dark Matter

Dark matter can be cold, i.e., at very small temperatures. This is what will happen if the mass is very large, so that the objects are almost non-relativistic. It can also happen if they are almost completely non-interacting.

1. *MACHOs*

These could be: brown dwarfs (dense like a white dwarf, except without light, i.e., no nuclear reactions happening inside them), Jupiters (would-be stars that have never ignited, like Jupiter itself, which are gaseous on the outside and not very dense), neutron stars (not emitting light), or massive black holes, etc. Of course, there are supermassive black holes at the center of every galaxy, but they represent only a small proportion of the galaxy's mass, and these are the most massive black holes there are, so they are excluded (general relativity simulations don't create overly many of them).

As we said, all of the MACHOs are excluded as a candidate for Dark Matter, both theoretically (it is hard to get enough of them to be created) and experimentally (we do not observe these MACHOs through lensing)

2. *WIMPs* (Weakly Interacting Massive Particles)

These are fundamental particles with mass that interact extremely weakly with other ones. They are the preferred explanation for Dark Matter.

3. *Axions*: more on them later

Hot Dark Matter

Dark matter can also be hot, i.e., thermalized at a temperature comparable to the CMBR's one. There is basically only one possibility, since having been in the past in thermal equilibrium with the CMBR is a condition, which pretty much restricts it to be a neutrino (other species would decouple much sooner). More on them shortly.

Modifications of Gravity

Modifications of gravity (like MOND, though that is a modification of the Newton force law, not the gravity law, or $f(R)$, or massive gravity, etc.) generally don't satisfy all of the constraints from the experimental evidences. It is very difficult to satisfy them all, since they appear at different scales (rotational curves for galaxy scale, then various at galaxy cluster scale, then CMBR scale), and a modification of gravity has generically a given distance scale at which it sets in.

Neutrino Hot Dark Matter

Neutrinos being the dark matter leads to the astrophysics constraint (see [5])

$$\sum_i m_{\nu_i} \lesssim 0.3\,\text{eV}; \quad \text{(Planck)}; \quad (0.17\,\text{eV} \quad \text{combined}). \tag{4.17}$$

Direct particle physics searches for them lead to the constraints (from the Particle Data Group 2017, http://pdg.lbl.gov/2017/listings/rpp2017-list-neutrino-prop.pdf)

$$\begin{aligned} m_{\nu_e} &< 2\,\text{eV} \\ m_{\nu_\mu} &< 190\,\text{KeV} \\ m_{\nu_\tau} &< 18\,\text{MeV}. \end{aligned} \tag{4.18}$$

On the other hand, the tightest constraints come from neutrino oscillations. The observed neutrino oscillation depends only on the Δm^2 of the two neutrino species. For the species 1, 2, 3 ordered in terms of increasing mass, we find (from the Particle Data Group 2018, http://pdg.lbl.gov/2018/reviews/rpp2018-rev-neutrino-mixing.pdf)

$$\Delta m_{12}^2 = 7.4 \times 10^{-5}\,\text{eV}^2; \quad \Delta m_{31}^2 = 2.5 \times 10^{-3}\,\text{eV}^2. \tag{4.19}$$

This in turn leads to a bound on the highest neutrino mass eigenstate of

$$m_{\nu_{\text{highest}}} < 0.05\,\text{eV}\ , \tag{4.20}$$

which is the tightest constraint we have.

But the Hot Dark Matter scenario is excluded by the CMBR at low angular momentum l. Indeed, in Hot Dark Matter the fluctuation spectrum $\delta\rho/\rho$ or $\delta T/T$ at low l would drop to zero, for a linear increase with l. But even with the experimental error bars, it is clear that in reality the CMBR goes to a constant, so the main contribution to Dark Matter is Cold, though there could also be a small Hot component.

4.3 WIMP Dark Matter Candidates

Since Cold Dark Matter of the WIMP variety is the choice preferred by experiments, it is worth saying what are the main candidates.

Within supersymmetry, still one of the preferred models of high energy completion in the Standard Model, the clear Dark Matter candidate is the *Lightest Supersymmetry Particle, or LSP*. We don't need to know the details of the supersymmetric model, just that there is a particle of smallest mass that therefore cannot decay (is stable), is relatively light, so it can act as Dark Matter, and has almost no interactions with the other particles.

In supersymmetry, we have the MSSM (Minimally Supersymmetric Standard Model) or its extensions, forming the *visible sector*. We also have a *hidden sector*, which is a strongly coupled sector that breaks supersymmetry through some non-perturbative mechanism (since we know that perturbative mechanisms are excluded by experiment in the Standard Model, and there is no right nonperturbative coupling in the MSSM; in the hidden sector there can be). Finally, we have *messenger fields*, or mediators, that carry the supersymmetry breaking from the hidden sector to the visible one (MSSM). This mediation of supersymmetry breaking is usually made by gauge fields (gauge mediation) or gravity (gravity mediation).

So the only exception to the LSP being Dark Matter comes from *gravity mediated scenarios*, when the *gravitino* (superpartner of the graviton) is the LSP. This is generically inconsistent with cosmology. In this case, amplitudes for annihilation go like G_N instead of $\sqrt{G_N}$, so the annihilation rates go like

$$\Gamma \propto \mathcal{A}^2 \propto G_N^2 \ , \tag{4.21}$$

which is too small for the annihilation rate to be useful for giving gravitino decay. But in this case, the number of gravitinos is about one percent of the number of photons (which is similar to the case of the neutrinos), from which (if the gravitino is the Dark Matter) we can derive a bound on the mass of the gravitinos,

$$m_{\text{gravitino}} \lesssim 100\,\text{eV}. \tag{4.22}$$

But this is too small to make sense in the gravity mediation scenarios, so in this case the cosmology contradicts the simplest gravity mediation scenarios.

If however the gravitino is not the LSP and can decay, giving

$$\Gamma \propto \mathcal{A}^2 \propto G_N \ , \tag{4.23}$$

then one finds (like for any decaying particle) from the Big Bang Nucleosynthesis constraints that they need to decay before it, which translates into

$$m_{\text{gravitino}} > 10\,\text{TeV}. \tag{4.24}$$

This in turn means a supersymmetry breaking scale

$$M_s > 10^{11}\,\text{GeV} \ , \tag{4.25}$$

since

$$m_g \simeq \sqrt{G_N} M_s^2. \tag{4.26}$$

This is a very high scale, and it means that the gravitino is not Dark Matter in this scenario.

Axions

Another possibility is for Dark Matter to be coming from *axions*.

In the Standard Model, there is a term possible to be added to the Lagrangian of QCD, the theta term

$$\int d^4x \theta\, \text{Tr}[F_{\mu\nu}\tilde{F}^{\mu\nu}] \ , \tag{4.27}$$

where $\tilde{F}^{\mu\nu} = \frac{1}{2}\epsilon^{\mu\nu\rho\sigma}F_{\rho\sigma}$ is the Maxwell dual to the YM field.

But experimentally, it is found to be extremely small ($\theta \ll 1$), which can be enforced by the existence of a symmetry, called Peccei–Quinn symmetry, that would

put $\theta = 0$. But if we spontaneously break the symmetry, the Goldstone boson for the broken Peccei–Quinn symmetry is called an *axion* and is identified with a varying $\theta(x)$. An axion could have a mass

$$m_{\text{axion}} \sim 10^{-3} - 10^{-6}\,\text{eV} \tag{4.28}$$

and would give a good candidate for Dark Matter. These axions appear easily within string theory, so they can be relevant for cosmology.

4.4 The ΛCDM Model

In conclusion, we see that the *Standard Model of Cosmology*, that best fits data, is a ΛCDM (cosmological constant plus dark matter) model, with proportions:

4% baryons, 23% cold dark matter and 73% cosmological constant.

We say cosmological constant, but really, what is sure is we have an accelerated Universe (which as we saw requires an equation of state with $w < -1/3$), but cosmological constant ($w = -1$) fits best the data. In principle it need not be a cosmological constant, but rather a *dark energy* component, which could have a time-varying w. However, there is no truly generic constraint on the time dependence of w, one has to make some assumption, and the Planck data constrains various scenarios.

Important Concepts to Remember

- The first method to observe dark matter was the Coma Cluster of galaxies, measuring statistically the virialized velocities of individual galaxies, and seeing that $v^2 \gg G_N M_{\text{visible}}/r$.
- The most classic test however is the rotational curves of galaxies, i.e., test stars rotating around the galaxy. Since $v = \sqrt{G_N M(r)/r}$ and we observe $v \simeq$ constant, it means that there is dark matter with $M(r) \propto r$ outside the visible one.
- A crucial test eliminating most alternatives to dark matter is the Bullet Cluster of galaxies, composed mostly of gas. Two clusters have collided, and the stars are ahead, but most of the gas is in between. However, the dark matter is still around the stars, so the visible and dark matter are in different regions.
- Gravitational lensing means focusing of light from a distant source by a mass distribution in front of it (with respect to us). It allows us to see dark matter distributions in front of some distant sources.
- Big Bang Nucleosynthesis is sensitive to Ω_{baryons}.
- The CMBR is sensitive to Ω_{total}, but also to Ω_{baryons}, allowing us to measure the amount of dark matter.
- Hot Dark Matter (usually neutrinos) is excluded, since the fluctuation spectrum at small l would drop to zero fast.
- MACHO cold dark matter, i.e., brown dwarfs, etc., cannot explain dark matter. We need WIMPs or axions for Cold Dark Matter to do so.

- The Lightest Supersymmetric Particle (LSP) is a good WIMP, however gravitinos are not.
- Axions are also a possibility for WIMPs.
- The standard ΛCDM model has 73% dark energy, 23% cold dark matter and 4% baryons.

Further reading: See the reviews [6, 7]. Also [8] for the Bullet Cluster.

Exercises

(1) (MOND theory) An alternative explanation for the rotational curves of galaxies is in terms of a modification of the Newton's force law at low acceleration, from $a = F/m$ to $a = \sqrt{a_0 F/m}$. Show that the latter leads to the observed $v = \text{const.}(r)$. Choose an interpolating function between the 2 regimes and find the exact $V(r)$ profile.
(2) Is it possible for the gravitational lensing effects to cancel against one another (to have false negatives: that we see nothing in the sky, yet there is lensing)? Why?
(3) Calculate the time evolution of the total energy density $\rho(t)$ in the ΛCDM model with the exact proportions Ω_i in the text.
(4) Consider noninteracting dark matter particles of mass $m \ll k_B T_{\text{DM}}$, thus relativistic, and thermalized at temperature T_{DM} of the order of, but not equal to, the CMBR temperature of 3.75 K. If dark matter is *entirely* composed of them, calculate their number density.

Chapter 5
The Early Universe and Its Thermal History

In this chapter, I will describe the thermal history of the Early Universe, i.e., the evolution of the particle species in it. After describing the Hot Big Bang scenario, and why it is required, we will calculate the thermodynamics of various particle species in the Universe. The evolution of the particle species during the radiation dominated era will be then described, and finally we will look in detail at neutrino decoupling.

5.1 Defining the Hot Big Bang

The Universe starts off in a *hot, equilibrium* state. It is an *equilibrium* state, since at $t = 1$ s from the Big Bang, the characteristic reaction time for e^+e^- annihilation, $e^+e^- \leftrightarrow 2\gamma$, is $\tau = 10^{-17}$ s, so there is more than enough time to reach equilibrium in the annihilation reaction.

In order to characterize the Universe, we will use the quantities: entropy (per baryon)

$$S_1 \simeq \frac{N_\gamma}{N_{\text{baryon}}} , \tag{5.1}$$

(we will see later in the chapter that the relation is correct up to a numerical factor of order one) and normalized (electron) lepton number

$$L_e = \frac{(N_{e^-} - N_{e^+}) + (N_\nu - N_{\bar{\nu}})}{N_{\text{baryon}}}. \tag{5.2}$$

The next assumption to check is whether the Universe starts off in a hot state. We will do that historically, following the various proposals that were considered before the Hot Universe:

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_5

(1) **Only Neutrons Initially**

Looking at the Universe, it seems like it is a cold place, and mostly made up of protons and neutrons, plus electrons. It is also electrically neutral, so the simplest assumption is that it starts in a state with only neutrons n. Since n interacts with p (protons), it is hoped that we can create the necessary protons.

The initial quantum numbers of the Universe are, therefore, $S_1 = 0$ (cold Universe) and $L_e = 0$.

The neutrons (outside a nucleus) can decay, with a half time $\tau \sim 1000\,\mathrm{s}$, and they start a reacting chain

$$\begin{aligned} n &\to p + e^- + \bar{\nu} \\ p + n &\to D + \gamma \\ D + D &\to T + p \\ &\to {}^3He + n \\ {}^3He + n &\to T + p \\ T + D &\to {}^4He + n. \end{aligned} \tag{5.3}$$

Specifically, the neutrons decay into protons, which then react with the protons to make deuterium. Two deuterium make tritium and more protons, and the tritium and deuterium make helium 4. Thus the chain starts with neutrons, and ends up with helium 4, but no protons (all the protons react with n and decay). That means we don't produce protons, i.e., hydrogen nuclei. This is bad, since the Universe is mostly made up of hydrogen.

(2) **Only Protons and Electrons in Equal Numbers**

The next possibility is to *start* with protons p's, but then in order for the Universe to be electrically neutral, we must have an equal number of electrons e^-s, $N_p = N_{e^-}$. In that case $S_1 = 0$ (cold Universe), $N_\gamma = 0$, but $L_e = 1$.

However, in this case, we start off with the reactions

$$\begin{aligned} p + e^- &\to n + \nu \\ p + n &\to D + \gamma\,, \end{aligned} \tag{5.4}$$

and from then on, the reaction chain follows in the same way, ending up at ^{4}He. So in this case as well, we consume all protons, and end up with only Helium 4. Again, this is bad, since in the end we have no hydrogen (even though we started with ionized hydrogen atoms, i.e., protons and electrons in equal amounts). But the point is that protons and neutrons interact until there is nothing left.

(3) **Protons, Electrons and Neutrinos in Equal Numbers**

Next in complexity is to assume that we have also neutrinos in equal numbers with the protons and electrons, which doesn't change the neutrality of the Universe. In this case $S_1 = 0$ and $L_e = 2$.

Now the starting point of the reaction chain,

$$p + e^- \rightarrow n + \nu \tag{5.5}$$

is forbidden because of Pauli's exclusion principle. The neutrinos are fermions, and there are already neutrinos in the state, occupying the same states.

Since the reaction chain cannot start, the protons and electrons can bind to form hydrogen atoms. So H is produced, but nothing else, so that is not good either.

All of these solutions involved a cold Universe ($S = 0$).

(4) **The Hot Universe**

But in the 1940s, George Gamow finally came up with the hypothesis of the Hot Universe, with $S_1 \gg 1$ and $L_e \sim 1$, that is, we have many photons for each baryon, and equally electrons and neutrinos, comparable with the baryons.

Now, due to the many photons present, the reaction

$$p + n \rightarrow D + \gamma \tag{5.6}$$

is not possible, since it occurs equally in the opposite direction, with the photons destroying the Deuterium atoms. That means that the reaction chain (cycle) cannot continue beyond the formation of protons (Hydrogen nuclei).

But that is fine! Since that means that there is mostly hydrogen in the Universe, and there is only 20, 30% Helium. With the drop in temperature, we can create *some* Deuterium, during Big Bang Nucleosynthesis, as will be showed in the next chapter. In terms of abundances of isotopes, the ones with more neutrons predominate, since we have free neutrons for a long time, and we have high temperature.

5.2 Equilibrium Thermodynamics

We have seen that the Universe is hot and in equilibrium, so we now describe equilibrium thermodynamics, within relativistic theory.

We denote the Fermi–Dirac or Bose–Einstein distributions by

$$f(\vec{p}) = \frac{1}{e^{\frac{E-\mu}{k_B T}} \pm 1}. \tag{5.7}$$

For a relativistic particle with degeneracy g, the number density is

$$n = \frac{g}{(2\pi)^3} \int f(\vec{p}) d^3 p. \tag{5.8}$$

The energy density is found by introducing an energy $E(\vec{p})$ inside the integral, i.e.,

$$\rho = \frac{g}{(2\pi)^3}\int E(\vec{p}) f(\vec{p}) d^3p. \tag{5.9}$$

The pressure is found from relativistic kinetic theory,

$$P = \frac{g}{(2\pi)^3}\int \frac{|\vec{p}|^2}{3E} f(\vec{p}) d^3p. \tag{5.10}$$

In the kinetic theory of gases, the pressure is found as momentum transfer in the x direction, perpendicular to the surface, times the velocity, i.e., $\langle p_x^2\rangle/m = \langle p^2\rangle/(3m)$. In the relativistic case, $m \to E$, and $|\vec{p}| = \sqrt{E^2 - m^2}$.

Writing $d^3p = 4\pi p^2 dp$, and using $p^2 = E^2 - m^2$, we have $d^3p = 4\pi E dE \sqrt{E^2 - m^2}$. Replacing in the integrals, we obtain

$$\begin{aligned} n &= \frac{g}{2\pi^2}\int_m^\infty \frac{(E^2-m^2)^{1/2}}{e^{\frac{E-\mu}{k_BT}} \pm 1} E dE \\ \rho &= \frac{g}{2\pi^2}\int_m^\infty \frac{(E^2-m^2)^{1/2}E^2}{e^{\frac{E-\mu}{k_BT}} \pm 1} dE \\ P &= \frac{g}{6\pi^2}\int_m^\infty \frac{(E^2-m^2)^{3/2}}{e^{\frac{E-\mu}{k_BT}} \pm 1} dE. \end{aligned} \tag{5.11}$$

We are however not so interested in the intermediate region of energies described by the general formulas above, but rather in the extreme regions: ultrarelativistic and non-relativistic.

Relativistic Species

Consider now (ultra)relativistic species, i.e., $T \gg m, \mu$. In that case, we obtain

$$\begin{aligned} n &= \frac{\zeta(3)}{\pi^2} g T^3; \quad \text{Bose} \\ &= \frac{3}{4}\frac{\zeta(3)}{\pi^2} g T^3; \quad \text{Fermi.} \end{aligned} \tag{5.12}$$

Note that in the relativistic case, the integrals can be taken from 0 instead of $m \ll T$, and then by rescaling the energies by T, we obtain the scaling of n with gT^3, with a numerical prefactor given by the adimensional integral. Similarly, for the energy density we obtain

$$\begin{aligned} \rho &= \frac{\pi^2}{30} g T^4; \quad \text{Bose} \\ &= \frac{\pi^2}{30}\frac{7}{8} g T^4; \quad \text{Fermi.} \end{aligned} \tag{5.13}$$

In the case of the pressure, in both cases we obtain

$$P = \frac{\rho}{3}\,, \tag{5.14}$$

as we should. Note that the Bose case is exemplified by photons (γ), and the Fermi case is exemplified by neutrinos (ν).

We can also calculate the average energy per particle for the relativistic species

$$\begin{aligned}\langle E_1\rangle \equiv \frac{\rho}{n} &= \frac{\pi^4}{30\zeta(3)}T \simeq 2.7T; \quad \text{Bose} \\ &= \frac{7}{6}\frac{\pi^4}{30\zeta(3)} \simeq 3.15T; \quad \text{Fermi.}\end{aligned} \tag{5.15}$$

However, in the relativistic reactions, usually one has at least two species, particles and their antiparticles, and they are in equilibrium, with $\mu_+ = -\mu_- \equiv \mu$. Substituting in the general formulas, we obtain

$$\begin{aligned}n_+ - n_- &= \frac{gT^3}{6\pi^2}\left[\pi^2\frac{\mu}{T} + \left(\frac{\mu}{T}\right)^3\right]; \quad T \gg m \\ &= 2g\left(\frac{mk_BT}{2\pi}\right)^{3/2}\sinh\frac{\mu}{k_BT}e^{-\frac{m}{k_BT}}; \quad T \ll m.\end{aligned} \tag{5.16}$$

In the second line, we have also considered the nonrelativistic case, in which case $E \simeq m$ and the FD and BE distributions become Maxwell, which moreover (since $E \simeq m$) gets outside the integral, giving the overall $e^{-\frac{m}{T}}$ factor.

Nonrelativistic Species

In the nonrelativistic case $T \ll m, \mu$, we obtain (in the same way as for $n_+ - n_-$ above)

$$n = g\left(\frac{mk_BT}{2\pi\hbar^2}\right)^{3/2} e^{-\frac{m-\mu}{k_BT}}. \tag{5.17}$$

For the energy density, we have an extra $E \simeq m$ inside the integral, so we have

$$\rho = mn. \tag{5.18}$$

Insider the integral for the pressure, we have an extra $(E^2 - m^2)/(3E)$, which finally gives the usual ideal gas law

$$P = nk_BT \ll \rho. \tag{5.19}$$

5.3 Thermodynamics of the Universe

In order to consider what happens in the expanding Universe, we need to consider the case of several species of relativistic particles, each in equilibrium at different temperature T_i. This is possible, since we can have decoupled systems that only expand with the Universe, but keep their thermal distribution. In this case, we define

$$\begin{aligned} \rho_R &= \frac{\pi^2}{30} g_* T^4 \\ P_R &= \frac{\rho_R}{3} = \frac{\pi^2}{90} g_* T^4 \,, \end{aligned} \tag{5.20}$$

where T is the temperature of photons. This is really a definition of g_*. By summing over the various components, we have

$$g_* = \sum_{i=\text{bosons}} g_i \left(\frac{T_i}{T}\right)^4 + \sum_{i=\text{fermions}} g_i \left(\frac{T_i}{T}\right)^4 \,, \tag{5.21}$$

and this is the *number of effectively massless degrees of freedom* (or, ultrarelativistic).

The expansion of the Universe is *adiabatic*, which means that the entropy per *comoving* volume is conserved. We write

$$T dS = d(\rho V) + P dV = d[(\rho + P)V] - V dP. \tag{5.22}$$

But from the general relativistic formulas (5.20), we obtain

$$\frac{dP}{dT} = \frac{4}{3}\frac{\rho}{T} = \frac{\rho + P}{T} \,, \tag{5.23}$$

which allows us to rewrite

$$\begin{aligned} T dS = d[(\rho + P)V] - \frac{V dT}{T}(\rho + P) \Rightarrow \\ dS = d\left[\frac{(\rho + P)V}{T} + \text{const.}\right] \end{aligned} \tag{5.24}$$

That means that the entropy per comoving volume is

$$s = \frac{S}{V} = \frac{\rho + P}{T} = \frac{4}{3}\frac{\rho}{T} = \frac{2\pi^2}{45} g_{*s} T^3 \,, \tag{5.25}$$

where in the last equality we have written a formula for several species. Just that we need to do that *after* dividing by T for each individual species, so that the g_* that appears is not the one in ρ_R, which was defined with factors of T^4, but rather is one with factors of T^3, namely

$$g_{*s} = \sum_{i=\text{bosons}} g_i \left(\frac{T_i}{T}\right)^3 + \sum_{i=\text{fermions}} \frac{7}{8} g_i \left(\frac{T_i}{T}\right)^3 . \tag{5.26}$$

On the other hand, specializing for photons γ, the number density is

$$n_\gamma = \frac{\zeta(3)}{\pi^2} g_\gamma T^3 , \tag{5.27}$$

where $g_\gamma = 2$, for the two photon polarizations. Combining with (5.25) and substituting the numerical values for the constants, we obtain

$$s = 1.80 g_{*s} n_\gamma , \tag{5.28}$$

or

$$S_1 = \frac{S}{N_{\text{baryons}}} = \frac{s}{n_{\text{baryons}}} = 1.80 g_{*s} \frac{N_\gamma}{N_{\text{baryons}}} , \tag{5.29}$$

justifying the formula from the beginning of the chapter.

Moreover, putting numbers, we get the formula

$$n_\gamma(\text{photons per } cm^3) \simeq 20 T^3 ({}^0 K). \tag{5.30}$$

Physical Values

We can measure that

$$\rho_{\text{rad,CMBR}} \simeq 30 \rho_{\text{rad,all sources}} , \tag{5.31}$$

so the main part of the radiation in the Universe is in the CMBR. We measure the temperature of the CMBR as

$$T_{\text{CMBR}} \simeq 2.7\,\text{K}. \tag{5.32}$$

According to the formula above, the density of photons in the Universe is

$$n_{\gamma,0} \simeq 20 \times (2.7)^3 \simeq 410\,\text{cm}^{-3}. \tag{5.33}$$

In turn, either by using the Stefan-Boltzmann law, or by the relation between ρ_γ and n_γ from the relativistic thermodynamics, we find

$$\rho_{\gamma,0} = \sigma T^4 \simeq 4.64 \times 10^{-34}\,\text{g/cm}^3. \tag{5.34}$$

Since

$$\frac{\rho_{\text{rad},0}}{\rho_{\gamma,0}} = 1 + \frac{7}{8} 3 \left(\frac{T_\nu}{T_\gamma}\right)^3 , \tag{5.35}$$

where the factor of 3 comes from the 3 species of neutrinos, and the temperature of neutrinos will be estimated at the end of the chapter, we find

$$\rho_{\text{rad},0} \simeq 7.8 \times 10^{-34}\,\text{g/cm}^3. \tag{5.36}$$

The number density of baryons can be estimated (how? to do in exercise) and is approximately equal to the number of protons,

$$n_{\text{baryons}} \simeq n_p \simeq 10^{-6}\,\text{cm}^{-3}\ , \tag{5.37}$$

from which we find (multiplying with the mass of the proton) the mass density of baryons,

$$\rho_{m,0} \simeq 10^{-30}\,\text{g/cm}^3. \tag{5.38}$$

Then we have the ratio

$$\frac{\rho_{m,0}}{\rho_{\gamma,0}} \simeq 2000. \tag{5.39}$$

We also find (putting in the most precise value for the result)

$$\frac{N_\gamma}{N_{\text{baryons}}} = \frac{n_\gamma}{n_{\text{baryons}}} \simeq 1.66 \times 10^9. \tag{5.40}$$

But we know that the matter and radiation scale respectively as $\rho_m \propto R^{-3}$, $\rho_\gamma \propto R^{-4}$, therefore we have equality between the matter and radiation energy densities, $\rho_m \simeq \rho_\gamma$ at

$$\frac{R_0}{R} \simeq 2000\ , \tag{5.41}$$

which means at the temperature (the temperature scales inversely with the radius)

$$T \simeq 2000 T_0 \simeq 6000\,\text{K}\ , \tag{5.42}$$

which incidentally, is about the temperature of the Sun. Also, since the scale factor R scales as $1/(1+z)$ (z is the redshift of the light coming from this time in the past), we have matter radiation equality at

$$1 + z_{eq} = \frac{R_0}{R} \Rightarrow z_{eq} \simeq 2000. \tag{5.43}$$

or, since $R \simeq t^{1/2}$ during the matter dominated (M.D.) era, we have

$$t_{eq} \simeq 1500 \text{ years} \simeq 5 \times 10^{10}\,\text{s}. \tag{5.44}$$

5.4 Particle Species in the Radiation Dominated (R.D). Era

Around the time of matter-radiation equality happens also recombination, the time when atoms form and the CMBR last scatters. After that, there are no more thermal events in the Universe, the only interesting events being the formation of galaxies and later stars.

So all of the thermal history of the Universe happens in the radiation dominated (R.D.) era. We have given the relation of $\rho(t)$ in the R.D. era before; writing it in numbers, we get

$$\rho_\gamma(g/cm^3) \simeq \frac{10^6}{t(s)^2}. \tag{5.45}$$

Another useful formula is the transformation between temperature (expressed in Kelvin or MeV) and time,

$$T(^0\,K) = \frac{10^{10}}{t(s)^{1/2}} \Rightarrow T(MeV) \simeq \frac{1\,\text{MeV}}{t(s)^{1/2}}. \tag{5.46}$$

As we said at the beginning of the chapter, at times $t < 1\,\text{s}$, i.e., at temperatures $T > 10^{10}\,\text{K} = 1\,\text{MeV}$, the Universe is in equilibrium, since the reactions like $e^+e^- \leftrightarrow 2\gamma$ or $e^+e^- \leftrightarrow \nu_e\bar{\nu}_e$ occur very quickly, in a time much less than the age of the Universe t.

Before this time, there are many other particles in equilibrium, not just $e^\pm$, γ. The number of photons divided by the number of baryons was found in (5.40) to be about 10^9.

The entropy per baryon S_1 is conserved in the expansion of the Universe (since $S_1 = s/n_{\text{baryon}}$ and both the entropy per comoving volume s, as a result of adiabaticity, as we said, and the number density of baryons per comoving volume n_{baryon}, are constant). But we can rewrite (5.29) as

$$\frac{N_{\text{baryon}}}{N_\gamma} = \frac{n_{\text{baryon}}}{n_\gamma} \simeq 1.80 g_{*s} \frac{n_{\text{baryon}}}{s}\,, \tag{5.47}$$

and since $n_{\text{baryon}}/s = 1/S_1$ is conserved, it follows that $N_{\text{baryon}}/N_\gamma$ changes only when g_{*s} changes.

But g_{*s} changes due to particle-antiparticle annihilations ending when the temperature T drops below the rest mass of the particle E_0 (as a result, annihilations only happen with exponentially small probability, de facto stopping). In that case, the particle-antiparticle pairs disappear, they dump their entropy into the photons, increasing n_γ. The excess of particles over antiparticles becomes then nonrelativistic.

Using the relation between the energy in MeV, for the rest energy of the particle, and the temperature in Kelvin,

$$E(MeV) = k_B T \simeq 10^{-10} T(^0K)\,, \tag{5.48}$$

(so that $10^{10}\,\mathrm{K} = 1\,\mathrm{MeV}$), we find the temperatures of various particles (when the annihilation stops and the remaining particles become nonrelativistic)

$$\begin{aligned} T_e &\simeq 5 \times 10^9\,\mathrm{K} \quad \text{for} \quad m_e \simeq 0.5\,\mathrm{MeV} \\ T_\mu &\simeq 10^{12}\,\mathrm{K} \quad \text{for} \quad m_\mu \simeq 100\,\mathrm{MeV} \\ T_p &\simeq 10^{13}\,\mathrm{K} \quad \text{for} \quad m_p \simeq 1\,\mathrm{GeV}. \end{aligned} \tag{5.49}$$

Thus, for $T < T_e$, electrons decouple from radiation, for $T < T_\mu$, muons do, and for $T < T_p$, protons do as well.

That also tells us how it came about that we have such high $N_\gamma/N_{\mathrm{baryons}}$: the various baryons annihilate, and their numbers and entropies are changed into photons. Going back in time, all the N_γ would eventually correspond to $N_{\mathrm{baryon}} = N_{\mathrm{anti-baryon}}$. Indeed, at equilibrium for these baryons, we have $N_{\mathrm{baryons}} \simeq \bar{N}_{\mathrm{anti-baryons}} \simeq N_\gamma$. The (small) amount of baryons we see in the Universe today is only the excess of initial baryons over antibaryons, which therefore cannot annihilate (without it, the previous approximate equality would be exact, $\Delta N_{\mathrm{baryons}} = \bar{N}_{\mathrm{anti-baryons}}$. Then the initial baryon to antibaryon ratio is

$$\frac{N_{\mathrm{baryon,in.}}}{\bar{N}_{\mathrm{antibaryon,in.}}} \simeq \left(\frac{n_\gamma + n_{\mathrm{baryon}}}{n_\gamma}\right)_0 \simeq \frac{10^9 + 1}{10^9}\,, \tag{5.50}$$

i.e., the initial baryon asymmetry was $1 : 10^9$ only.

Recombination

Recombination is the last scattering of photons, thus the time atoms form, without being dissociated by the scattering of photons. After it, the Universe becomes transparent to photons.

It occurs roughly when the mean free path of photons,

$$l_f \simeq \frac{1}{n\sigma} \simeq 10^2[t(s)]^{3/2}\,m\,, \tag{5.51}$$

equals the horizon size $l_{\mathrm{hor}} \simeq ct$ (during the M.D. era, it is a factor of 2 larger, during the R.D. era a factor of 3, but in any case a factor of order one away from ct). Here n is the number density of photons, which during the M.D. era decays as $n = 10^{23}t^{-3/2}$, and σ is their cross section (now, $\sigma_0 \simeq 10^{-24}\,\mathrm{cm}^{-2}$). In the last equality we have substituted these values for n and σ.

The time when the mean free path equals the horizon size (thus recombination happens, since the photons cannot interact anymore, not even as they travel the horizon distance) is

$$t^{1/2} = \frac{c}{10^2} \Rightarrow t = 10^{13}\,\mathrm{s} \simeq 3 \times 10^5 years, \tag{5.52}$$

This corresponds to a temperature $T \simeq 4000\,\mathrm{K}$. A more precise calculation will be carried out in the next chapter.

Partial Summary

So we have $T_\gamma \simeq 4 \times 10^3$ K for recombination, after which we have $T_e \simeq 5 \times 10^9$ K, $T_\mu \simeq 10^{12}$ K and $T_p \simeq 10^{13}$ K. Using the formula $T(K) = 10^{10}/[\mathrm{t(s)}]^{1/2}$, we find also

$$t_e \simeq 4\,\mathrm{s}, \quad t_\mu \simeq 10^{-4}\,\mathrm{s}; \quad t_p \simeq 10^{-6}\,\mathrm{s}. \tag{5.53}$$

5.5 Neutrino Decoupling

The last part of the thermal history of the Universe to be analyzed in this chapter is the decoupling of neutrinos. It is found from their equilibrium reactions with the corresponding leptons.

ν_e Decoupling

The relevant reaction is

$$e^+e^- \leftrightarrow \nu_e\bar{\nu}_e. \tag{5.54}$$

The cross section for this reaction is (as it occurs through the electroweak interaction that reduces to the 4-Fermi one)

$$\sigma(e^+e^- \leftrightarrow \nu_e\bar{\nu}_e) = \frac{g_F^2 E^2}{\hbar c^4} \simeq 10^{-63} T^2\ , \tag{5.55}$$

where in the last equality we have considered ultrarelativistic particles. Then, since $n \propto T^3$, the characteristic time (relaxation time) of the reaction is

$$\tau = \frac{1}{\sigma n c} \simeq \frac{10^{51}}{T(^0K)^5} s \propto t^{5/2}. \tag{5.56}$$

As in the case of recombination, the reaction stops when the characteristic time of the reaction becomes larger than the age of the Universe

$$t \simeq \frac{10^{20}}{T^2(K)} s \tag{5.57}$$

(note that the relaxation time increases faster than the horizon, so eventually the annihilation reaction creating neutrinos stops). The temperature when it stops is

$$T \simeq 10^{31/3}\,\mathrm{K} \simeq 2 \times 10^{10}\,\mathrm{K}\ , \tag{5.58}$$

corresponding to a time of

$$t_{\nu_e} \simeq \frac{10^{20}}{10^{62/3}}\,\mathrm{s} \simeq 0.25\,\mathrm{s}. \tag{5.59}$$

Note that this is slightly before the e^+e^- annihilation (which happens at 4s).

ν_μ Decoupling

The idea is similar, except now, since $t_\mu \simeq 10^{-4}$ s, at times after that, the remaining muons are nonrelativistic. It will indeed turn out that ν_μ decoupling happens after that, so in the reactions

$$\begin{aligned} \mu^+ + \mu^- &\leftrightarrow \nu_\mu + \bar{\nu}_\mu \\ \mu^+ &\leftrightarrow e^+ + \bar{\nu}_\mu + \nu_e \\ \mu^- &\leftrightarrow e^- + \nu_\mu + \bar{\nu}_e \ , \end{aligned} \tag{5.60}$$

the muons are nonrelativistic. That means that the reaction times are exponentially amplified (the reactions happen with exponentially smaller probabilities), and one finds a relaxation time of

$$\tau = 2 \times 10^{-6} e^{\frac{m_\mu c^2}{k_B T}} = 2 \times 10^{-6} \times e^{\frac{106}{T(MeV)}} \ , \tag{5.61}$$

since $m_\mu = 106\,\text{MeV}$. When the relaxation time is larger than the age of the Universe,

$$\tau > t = \frac{10^{20}}{T^2(K)} \ , \tag{5.62}$$

the reaction stops. This is a transcendental equation, which is solved by

$$T = 12\,\text{MeV} \simeq 10^{11}\,\text{K} \ , \tag{5.63}$$

corresponding to

$$t_{\nu_\mu} \simeq 0.01\,\text{s}. \tag{5.64}$$

ν_τ Decoupling

The same idea applies for the decoupling of ν_τ. Because of the exponential amplification of the relaxation time, one finds approximately the same temperature (and time) for the decoupling, despite now having $m_\tau \simeq 1778\,\text{MeV}$, so

$$\tau = 2 \times 10^{-6} e^{\frac{1778}{T(MeV)}} > t = \frac{10^{20}}{T^2(K)} \ , \tag{5.65}$$

results in about the same time,

$$t_{\nu_\tau} \simeq 0.01\,\text{s}. \tag{5.66}$$

Neutrino Temperature

At $t \simeq 0.01$ s, the muons are nonrelativistic and decoupled, but $\gamma, e^\pm$ and all the neutrinos are not. Therefore we find the ratios

$$\rho_\gamma : \rho_{e^\pm} : \rho_{\nu_e,\bar{\nu}_e} : \rho_{\nu_\mu,\bar{\nu}_\mu} : \rho_{\nu_\tau,\bar{\nu}_\tau} : \rho_{\mu^\pm} =$$
$$\rho_1 : \frac{7}{4}\rho_1 : \frac{7}{8}\rho_1 : \frac{7}{8}\rho_1 : \frac{7}{8}\rho_1 : 10^{-4}\rho_1. \tag{5.67}$$

The ratio of photons to neutrinos is due to the fact that there are two photons for each e^+ and e^- ($e^+e^- \leftrightarrow 2\gamma$), whereas one a fermion like the electron has $7/8$ of a boson's energy density. Both of these (photons and electrons) have two polarizations. The ratio of electrons to the various neutrinos is given by the fact that the electrons have two polarizations, whereas the neutrinos have only one, and otherwise the reactions give a one to one ratio in number density. Finally, the relation to $\rho_{\mu^\pm}$ is given by the factor $e^{-\frac{106}{12}} = 10^{-4}$ for the exponential drop in energy density in the nonrelativistic case.

After neutrino decoupling, we still continue to have $T_\nu = T_\gamma$, simply since both temperatures decay in the same way with the expansion of the Universe, until e^+e^- annihilation, happening at $t = 4\,\mathrm{s}$, when the energy of the electron-positron pairs is added to the energy of the photons. After that, the two temperatures, now different, scale with the Hubble expansion in the same way until now. Since as we saw

$$S_\gamma : S_{e^\pm} = 1 : \frac{7}{4}\,, \tag{5.68}$$

and the photon total entropy today (the entropy is conserved in the expansion, so this is the same as after e^+e^- annihilation) is given by the sum of the two entropies,

$$S_0(\gamma) = S_1(\gamma) + S_1(e^\pm) = \frac{11}{4}S_1(\gamma). \tag{5.69}$$

Since as we saw, the entropy of relativistic species goes like T^4, T^4 increases in the same proportion, and we find that the temperature of the photons today (which scaled after e^+e^- annihilation in the same way as the temperature of the neutrinos) is

$$T_{0,\gamma} = \left(\frac{11}{4}\right)^{1/3} T_{1,\gamma}\,, \tag{5.70}$$

whereas $S_0(\nu) = S_1(\nu)$, so $T_{0,\nu} = T_{1,\nu} = T_{1,\gamma}$ (at the e^+e^- annihilation the electrons and neutrinos still had the same temperature),

$$T_{0,\nu} = T_{0,\gamma}\left(\frac{4}{11}\right)^{1/3} \simeq 2\,\mathrm{K}. \tag{5.71}$$

Important Concepts to Remember

- The Universe starts off in a hot Universe in thermal equilibrium, with the entropy per particle (number of photons per baryon, up to a numerical constant) $S_1 \gg 1$ and the lepton number $L_e \sim 1$.

- For ultrarelativistic species, $n \propto T^3$ and $n_{\rm Fermi} = 3/4 n_{\rm Bose}$, and $\rho \propto T^4$ and $\rho_{\rm Fermi} = 7/8 \rho_{\rm Bose}$, and $P = \rho/3$.
- For nonrelativistic species, $\rho = mn$ and $P = nk_B T \ll \rho$, and $n \propto T^{3/2} e^{-\frac{m-\mu}{k_B T}}$.
- For relativistic species in the expanding Universe, $\rho_R = (\pi^2/30) g_* T^4$, with g_* the number of effectively massless degrees of freedom, and $s = S/V = (2\pi^2/45) g_{*,s} T^3$, and $s = 1.80 g_{*,s} n_\gamma$, or $S_1 = 1.80 g_{*,s} N_\gamma / N_{\rm baryons}$.
- The temperature of the CMBR now is $T \simeq 2.75 K$, and $N_\gamma / N_{\rm baryons} \sim 1.6 \times 10^9$ now.
- We have ρ_m / ρ_γ now of about 2000, so we are in the M.D. era, and equality with radiation was at $z_{\rm eq} \simeq 2000$.
- Since now $N_\gamma / N_{\rm baryons} \sim 10^9$, and photons were baryon-antibaryon pairs originally, the baryon asymmetry was originally of about 10^{-9}, i.e. $N_{\rm baryon} / N_{\rm antibaryon} \sim (10^9 + 1)/10^9$.
- Neutrinos decouple, but continue on a thermal distribution that scales as usual with $a(t)$.
- The temperature of the neutrinos now is $T_{\nu,0} \simeq T_{\gamma,0} (4/11)^{1/3} \simeq 2\,{\rm K}$.

Further reading: See Chap. 3 in [1] for more details.

Exercises

(1) Calculate g_* and $g_{*,s}$ now, from the known particles available.
(2) Estimate the total number of particles and the total number of baryons in the Universe.
(3) Calculate the total specific heat C_V in the Universe today.
(4) Given the thermal history of the Universe today, and the current matter composition, calculate the particle horizon as a function of H_0, the Hubble constant today.
(5) Calculate the number densities of ν_e, ν_μ, ν_τ in the Universe today, using the information in the text.

Chapter 6
Big Bang Nucleosynthesis and Recombination

In this chapter we will describe Big Bang Nucleosynthesis (BBN), the combination of fundamental protons and neutrons into nuclei, as well as calculate more precisely the temperature of recombination. We will start with that, since we will generalize it to the BBN case next. Then we will describe how BBN can be used to test ideas about the early Universe, since the primordial abundances, defined at BBN, depend on various cosmological parameter, and we can obtain experimental values for them.

6.1 Recombination Temperature

We consider the (chemical) reaction

$$p + e^- \leftrightarrow H + \gamma \,, \tag{6.1}$$

i.e., the recombination versus ionization of hydrogen. The reaction is in equilibrium (can occur both ways), and it has the ionization energy $I = 13.6\,\text{eV}$, the negative of the energy of the fundamental electron level in hydrogen. We will write the equation for the equilibrium, known as the *Saha equation*. It is based on the equation for the number density of a nonrelativistic species, found in the previous chapter,

$$n_i = g_i \left(\frac{2\pi m_i k_B T}{h^2}\right)^{3/2} e^{\frac{\mu_i - m_i}{k_B T}} = g_i \left(\frac{m_i k_B T}{2\pi\hbar^2}\right)^{3/2} e^{\frac{\mu_i - m_i}{k_B T}} . \tag{6.2}$$

Here g_i is the degeneracy factor of species i, and m_i its mass. Consider that in the reaction, the chemical potentials are the same on both sides of the equation, $\mu_p + \mu_e = \mu_H$, so by calculating the ratio $n_p n_e / n_H$, using that $m_p \simeq m_H$ ($m_p \simeq 1800 m_e$), we find

$$\frac{n_e n_p}{n_H} = \frac{g_e g_p}{g_H} \left(\frac{m_e k_B T}{2\pi\hbar^2}\right)^{3/2} e^{-\frac{I}{k_B T}} . \tag{6.3}$$

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_6

Since $g_e = g_p = 2$ (they are both spin 1/2 fermions) and $g_H = 4$ (there are 4 states) and putting $\hbar = 1$, we have

$$\frac{n_e n_p}{n_H} = \left(\frac{m_e k_B T}{2\pi}\right)^{3/2} e^{-\frac{I}{k_B T}}. \tag{6.4}$$

Then we define the ratio

$$X_e = \frac{n_p}{n} = \frac{n_p}{n_p + n_H} \tag{6.5}$$

and the inverse of the photon to baryon ratio,

$$\frac{n}{n\gamma} \equiv \eta \,, \tag{6.6}$$

which means that (since $n_p = n_e$ because of the neutrality of the Universe)

$$\frac{X_e^2}{1 - X_e} = \frac{n_p n_e}{(n_p + n_H) n_H} = \frac{n_p n_e}{n_H} \eta^{-1} \frac{1}{n_\gamma}. \tag{6.7}$$

Using the formula for relativistic bosons from the previous chapter, $n_\gamma = (\zeta(3)/\pi^2) g T^3 = (2\zeta/\pi^2) T^3$, since $g_\gamma = 2$, we find

$$\frac{X_e}{1 - X_e} = \frac{\sqrt{\pi}}{4\sqrt{2}\zeta(3)} \eta^{-1} \left(\frac{T}{m_e}\right)^{-3/2} e^{-\frac{I}{k_B T}}. \tag{6.8}$$

Using $n_0 = 10^{-6}\,\mathrm{cm}^{-3}$, $T_0 = 2.75\,\mathrm{K}$, we can find

$$\eta = 2.7 \times 10^{-8} (\Omega_B h^2) \,, \tag{6.9}$$

and replacing it in the above equation, we find

$$\frac{X_e^2}{1 - X_e} = 4 \times 10^{21} (\Omega_0 h_0^2)^{-1} T(K)^{-3/2} e^{-\frac{1.58 \times 10^5}{T(K)}}. \tag{6.10}$$

We can plot the resulting $X_e = X_e(T)$ and find that it drops to zero (we have total recombination of p's into H's) when the temperature drops through

$$T \simeq 4000\,\mathrm{K}. \tag{6.11}$$

That corresponds to $T \simeq 0.3\,\mathrm{eV}$, a redshift of

$$z_{\mathrm{rec}} \simeq 1300 \tag{6.12}$$

and a time of

$$t_{\rm rec} \simeq 10^{13}\,{\rm s}. \tag{6.13}$$

6.2 Big Bang Nucleosynthesis

It means the period of formation of nuclei, that gives origin at the various nuclear species observed today. To do that, we write a generalization of the Saha equation for the nuclear statistical equilibrium of nonrelativistic species. Starting with the same relation as before, now for a nuclear species $A(Z)$ (of atomic number A and charge Z)

$$n_A = g_A \left(\frac{2\pi m_A k_N T}{h^2}\right)^{3/2} e^{\frac{\mu_A - m_A}{k_B T}}. \tag{6.14}$$

The chemical reaction of nucleosynthesis is the formation of the nucleus from Z protons and $A - Z$ neutrons,

$$Z\ \ p's + (A - Z)\ \ n's \leftrightarrow A(Z). \tag{6.15}$$

Assuming moreover, as for recombination, that the chemical potentials on the two sides match,

$$\mu_A = Z\mu_p + (A - Z)\mu_n\ , \tag{6.16}$$

we find the equivalent of the Saha equation,

$$\frac{n_A}{(n_p)^Z (n_n)^{A-Z}} = g_A A^{3/2} 2^{-A} \left(\frac{h^2}{2\pi m_N k_B T}\right)^{\frac{3(A-1)}{2}} e^{\frac{I_A}{k_B T}}. \tag{6.17}$$

Here m_N is the nucleon mass, $m_p \simeq m_n = m_N$, and $m_A \simeq A m_N$, and I_A is the activation energy (equivalent to ionization). The total number of nucleons is

$$n_N = n_n + n_p + \sum_i (A n_A)_i\ , \tag{6.18}$$

and define the ratios

$$X_A \equiv \frac{A n_A}{n_N} \tag{6.19}$$

and X_n, X_p, such that $\sum_i X_i = 1$, we find for the fraction of species $A(Z)$

$$X_A = g_A \left[\zeta(3)^{A-1} \pi^{\frac{1-A}{2}} 2^{\frac{3A-5}{2}}\right] A^{5/2} \left(\frac{T}{m_N}\right)^{\frac{3(A-1)}{2}} \eta^{A-1} (X_p)^Z (X_n)^{A-Z} e^{\frac{I_A}{k_B T}}. \tag{6.20}$$

We can use this equation to find the fractions of various species produced during BBN, but for that we must find the initial conditions (the original numbers of fractions X_p and X_n when BBN starts).

Initial Conditions for BBN

Before BBN, the balance between protons and neutrons was maintained by weak interactions, through the reactions

$$\begin{aligned} n &\leftrightarrow p + e^- + \bar{\nu}_e \\ \nu + n &\leftrightarrow p + e^- \\ e^+ + n &\leftrightarrow p + \bar{\nu}_e. \end{aligned} \tag{6.21}$$

The first reaction is the decay of the neutron outside a nucleus, that at equilibrium can occur in both directions, the second is the reaction of a neutron with a neutrino to change into a proton and electron, and is characterized by the difference in energy

$$Q = m_n c^2 - m_p c^2 = 1.293\,\text{MeV}. \tag{6.22}$$

The reaction rate at high temperatures, $T \gg Q, m_e$, for the latter reaction is

$$\Gamma_{pe\to\nu n} \to G_F^2 T^5\ , \tag{6.23}$$

leading to the reaction characteristic time mentioned the previous chapter,

$$\tau(s) = \frac{1}{\Gamma} \simeq \frac{10^{51}}{T(K)^5}. \tag{6.24}$$

The reaction *freezes out* when τ becomes larger than the age of the Universe, $t(s) \sim 10^{20}/T^2(K)$. The resulting *freeze-out time* is almost the same, but a bit before, as ν decoupling,

$$T_* \simeq 0.8\,\text{MeV}. \tag{6.25}$$

At this time, both n and p are very nonrelativistic ($T \ll m_p, m_n$), and then the formula (6.2) applies, which leads to a ratio of n to p given by Q,

$$\left(\frac{n_n}{n_p}\right)_* = e^{-\frac{Q}{k_B T}} \simeq e^{-1.5} \simeq \frac{1}{5}. \Rightarrow X_{n,*} \simeq \left(\frac{n_n}{n_n + n_p}\right)_* \simeq \frac{1}{6} \tag{6.26}$$

More precisely, $X_{n,*} \simeq 0.1609$. After freeze-out, the only thing that still happens is that the neutron naturally decays to $p + e^- + \bar{\nu}_e$. Outside the nucleus, the decay occurs with a lifetime that is known surprisingly badly,

$$\tau_n = 10.5 \pm 0.2\,\text{min}. \tag{6.27}$$

The decay of the neutron happens until a time of about $t = 1 - 3\,\text{min}$, or about $T \simeq 0.3 - 0.1\,\text{MeV}$, when the temperature falls below the binding energy of Deuterium $D = {}^2H$, which is the first nucleus to form. Then we can use (6.20) with given n_n and n_p to find all other X_A's. The *boundary conditions*, n_n and n_p, are, at this time (D decoupling)

$$X_{n,D} \simeq \frac{1}{7}\,, \tag{6.28}$$

more precisely, $X_{n,D} \simeq 0.135$.

Around BBN, the number of effectively massless degrees of freedom, g_*, changes. Before BBN and until freeze-out $T = 0.8\,\text{MeV}$,

$$g_* \equiv \sum_{\text{bosons}} g_i \left(\frac{T_i}{T}\right)^4 + \frac{7}{8} \sum_{\text{fermions}} g_i \left(\frac{T_i}{T}\right)^4\,, \tag{6.29}$$

and all the massless particles have the same temperature. In the Standard Model, we have $e^{\pm}$, γ and ν that are effectively massless with the same temperature $T_i = T$, and the degeneracies are $g_\gamma = g_{e^\pm} = g_\nu = 2$ (but for the latter, considering ν and $\bar{\nu}$), which leads to

$$g_* = 2\left(1 + \frac{7}{8}(3+2)\right) = 10.75. \tag{6.30}$$

Here the 3 is from the ν_e (and $\bar{\nu}_e$), ν_μ (and $\bar{\nu}_\mu$), ν_τ (and $\bar{\nu}_\tau$), and the 2 is for e^+ and e^-.

But just before the nuclei start to form, more precisely around $T = 0.5\,\text{MeV}$, when we have e^+e^- annihilation (after which e^- become norelativistic), as we saw in the previous chapter, the entropy of the $e^\pm$ is transferred to the photons γ, which increases T_γ, and leads to a relation of the ν temperature with respect to $T \equiv T_\gamma$ of $T_\nu/T = (4/11)^{1/3}$, so that (there are no more e^+ and e^-, and the neutrinos have T_ν)

$$g_* = 2 + 2\frac{7}{8}3 \times \left(\frac{4}{11}\right)^{4/3} = 3.36. \tag{6.31}$$

At $t = 1$–$3\,\text{min}$, deuterium D starts to form, and after that, all the other nuclei form.

Once the deuterium starts to form, it starts a reaction chain

$$\begin{aligned} p + n &\to D + \gamma \\ D + D &\to {}^3H + p \\ D + D &\to {}^3He + n \\ {}^3H + D &\to {}^4He + n \\ {}^3He + n &\to {}^3H + p. \end{aligned} \tag{6.32}$$

This chain, as explained in the first chapter already, ends with all the neutrons trapped inside ^4He, the endpoint of the chain. That means that the *mass* abundance of ^4He is given by twice the abundance of neutrons, i.e.,

$$Y = 2\left(\frac{n_n}{n_n + n_p}\right)_* \simeq 0.27. \tag{6.33}$$

For each neutron trapped inside ^4He, there is a proton as well (^4He contains two neutrons and two protons), so the proton content drops as well.

Then at the end of the chain, there is some deuterium D and some ^3He left unburnt by the chain, specifically

$$\frac{D}{H}, \frac{^3He}{H} \sim 10^{-5} - 10^{-4}. \tag{6.34}$$

Moreover, using (6.20), we can see that also some ^7Li gets synthesized, namely

$$\frac{^7Li}{H} \sim 10^{-10} - 10^{-9}. \tag{6.35}$$

So we can plot the relative abundances of protons and neutrons and ^4He. The proton and neutron abundances start (at very small time) at 0.5 each, after which the n abundance drops, as the proton abundance raises correspondingly ($X_p = 1 - X_n$), since the mass of the neutron is slightly larger (so it decays more). As freeze-out ($T = 0.8\,\text{MeV}$), $X_n : X_p = 1 : 5$. Then until BBN, X_n drops very slowly, until $X_n : X_p = 1 : 6$. Then BBN starts, and all the X_n drops to zero, whereas $X_{^4He}$ rises (all neutrons are trapped in ^4He). In the same time, X_p drops by the same amount as X_n.

6.3 Dependence of BBN Primordial Abundances on Parameters

The primordial abundances turn out to depend on several parameters:

- the lifetime of the neutron $\tau_{1/2}(n)$. As we said, surprisingly, the neutron lifetime is known only with a large error. But since the reaction time for the $n + \nu \leftrightarrow p + e^-$ reaction, $\tau \simeq \frac{1}{\Gamma} \propto \tau_{1/2}(n)$. Since moreover the freeze-out occurs when $\tau \propto 1/T^5$ becomes larger than the age of the Universe, $t \propto 1/T^2$, it follows that the freeze-out temperature depends on the lifetime of the neutron as

$$T_* \propto [\tau_{1/2}(n)]^{1/3}. \tag{6.36}$$

Since the ^4He abundance Y depends drastically on T_*, it also depends on $\tau_{1/2}(n)$.

- the number of effectively massless degrees of freedom g_*. Since the energy density $\rho \propto g_* T^4$ (from the definition of g_*), and from the Friedmann equation $\rho \propto H^2$, it means that $H \propto g_*^{1/2} T^2$. But on the other hand, the age of the Universe is, up to a numerical constant of order one, $t \sim H^{-1}$, which means $t \sim g_*^{-1/2}$. But as above, $T_* \propto (\tau \sim t)^{-1/3}$, so
$$T_* \propto (g_*)^{1/6}. \tag{6.37}$$
We will shortly see that this gives a bound on the number of neutrino species N_ν.
- the ratio of the number of baryons to number of photons η. Indeed, we saw in (6.20) that $X_A \propto \eta^{A-1}$. Also, the amount of deuterium D and ^3He left unburnt after the neutrons are trapped into ^4He depends on η as $\propto \eta^{-n}$, with $n \sim 1-2$.
- also, the abundances depend on $\Omega_B h^2$, but from CMBR (that also depends on $\Omega_B h^2$) the latter is fixed rather well.

The various equations for the nuclear reactions, including the dependences above, is included into a numerical BBN code.

6.4 Observations and Comparison with BBN

We now turn to comparison with observations. We observe a deuterium abundance is
$$\frac{D}{H} \sim 2.78 \pm 0.4 \times 10^{-5}\,, \tag{6.38}$$
which from the BBN code implies that
$$\eta = 5.9 \pm 0.5 \times 10^{-10}. \tag{6.39}$$
Then, the observed ^3He abundance is
$$\left.\frac{^3He}{H}\right|_{\text{obs.}} \sim 1.1 \pm 0.2 \times 10^{-5}\,, \tag{6.40}$$
but it is not clear that this is a *primordial* abundance, since helium is also produced in stars, which burn hydrogen and create helium (like an H bomb). For the observed value of $\Omega_B h^2 = 0.0214$ (from the CMBR), the BBN code gives $^3He/H \simeq 1.04 \pm 0.06 \times 10^{-5}$, which agrees well with the observations, so perhaps all the observed amount above is of primordial origins.

^7Li Problem

From observations also, we see a Lithium 7 abundance of
$$\left.\frac{^7Li}{H}\right|_{\text{obs}} \simeq (2.07 + 0.15/-0.04) \times 10^{-10}. \tag{6.41}$$

But this is considerably less than the value obtained from the BBN code for $\Omega_B h^2 = 0.0214$, of about 3×10^{-10}. This is the Lithium 7 problem. There are ways to fix this, but at the expense of some other observable. For instance, the graph of the abundance of Lithium 7 against η has a minimum of 2×10^{-10}, but at an $\eta \sim 3 \times 10^{-10}$, which conflicts with the value extracted from the deuterium D abundance.

Number of Neutrino Species

From the BBN code, we can also constrain the g_* dependence, and the result is

$$N_\nu \geq 4 \,, \tag{6.42}$$

so at most four *light* neutrino species, whereas we know there are at least 3 light neutrino species (ν_e, ν_μ, ν_τ), $N_\nu \geq 3$.

Constraint on the Number of Massless Energy Modes

BBN constrains also, similarly to the g_* dependence, energy modes present before the creation of the Standard Model matter, usually graviton and inflaton modes, though others are also constrained. The constraint is

$$f_{BBN} = \frac{\rho_{\text{graviton}} + \rho_{\text{inflaton}} + \cdots}{\rho_{\text{matter}}} \leq 0.07. \tag{6.43}$$

But this constraint is easily satisfied in (most) inflationary models, as we will see later on in the book.

6.5 Summary of the Thermal History

We now summarize the thermal history of the Universe that we learned in this chapter and the previous one, adding a few events. The most common model of cosmology is based on inflation, which is usually at some high energy scale, a bit below the string scale E_s, itself usually a bit below the Planck scale E_{Pl}. The inflationary scale needs to be smaller than the string scale, since inflation is defined as an effective field theory involving a four dimensional scalar, and in the spirit of effective field theory, it needs to be valid below the cut-off, in this case string theory. There is also the a priori possibility that an ensemble of string modes gets excited at some large scale, like in the case of the only truly stringy cosmological model, the string gas model of Bradenberger and Vafa, which we will study in part III of the book. But otherwise, the string states are not excited, and we are in effective field theory.

- At $E \sim 10^{15}$ GeV, corresponding to $T \sim 10^{27}$ K, we can have a Grand Unified Theory (GUT) phase transition. From experimental evidence, there is a very strong probability that there is such a GUT scale, though it is not always so.
- There could be scales associated with other transitions, like a susy breaking scale M_{SSB}, etc.

- At $E = 300\,\text{GeV}$, corresponding to $T \sim 3 \times 10^{15}\,\text{K}$, we have the electroweak phase transition.
- At $E = m_p \simeq 900\,\text{MeV}$, the protons become nonrelativistic (protons and antiprotons annihilate).
- At $E \simeq 200\,\text{MeV}$ or a bit below, we have the QCD phase transition.
- At $E = m_\mu \simeq 100\,\text{MeV}$, the muons become nonrelativistic (muons and antimuons annihilate).
- At $E \simeq 10\,\text{MeV}$ ($T \simeq 10^{11}\,\text{K}$), both ν_μ and ν_τ decouple.
- At $E \simeq 1\,\text{MeV}$ ($T \simeq 10^{10}\,\text{K}$), ν_e decouple.
- At $E = 0.8\,\text{MeV}$, we have the freeze-out of the neutron/proton reaction.
- At $E = m_e = 0.5\,\text{MeV}$, the electrons become nonrelativistic, i.e., e^+e^- annihilation.
- Then, the BBN occurs, with the formation of light elements.
- The end of the R.D. era happens at $T \simeq 10^5\,\text{K}$, corresponding to a redshift of $z \sim 2 \times 10^4$, or a time of $t \sim 1.5 \times 10^3$ years = 6×10^{10} s.
- After it, we have M.D., with $T \propto t^{-2/3}$ ($\rho \propto 1/t^2 \propto 1/R^3 \propto T^3$).
- Recombination and the appearance of the CMBR (last scattering) happens at a temperature $T \sim 4000\,\text{K}$, corresponding to an energy $E \sim 0.5\,\text{eV}$, or a time $t \sim 5 \times 10^{12}\,\text{s} \sim 3 \times 10^5$ years, or a redshift of $z \simeq 1300$.

Important Concepts to Remember

- The recombination temperature (for atoms) is found from the Saha equation, when $X_e(T) = n_p/n$ drops to zero. One finds $T \simeq 4000\,\text{K}$, $z_{\text{rec}} \simeq 1400$ and $t_{\text{rec}} \sim 10^{13}\,\text{s}$.
- One writes down Saha equations for nucleons to form nuclear species $A(Z)$, for Big Bang Nucleosynthesis.
- Before BBN, the initial condition is as follows. The protons an neutrons are in equilibrium, with $(n_n/n_p) \simeq e^{-Q/k_B T}$ that freezes out at about 1/5. Then neutrons decay into protons a bit more, until the ratio is about 1/6.
- Just before BBN, g_* drops from 10.75 to 3.36.
- The ^4He to H ratio created in BBN is about 0.27, with D and ^3He of about $10^{-5} - 10^{-4}$, and ^7Li of about $10^{-10} - 10^{-9}$.
- Primordial abundances depend on: the neutron lifetime $\tau_{1/2}(n)$, the number of effectively massless degrees of freedom g_*, the number of photons per baryon $1/\eta$, $\Omega_B h^2$.
- We have a ^7Li problem: can't fit it together with the rest (only by itself).
- We constrain the number of neutrino species from BBN.
- We constrain the number of massless energy modes, like gravitons and inflaton modes, to $\rho_{\text{modes}}/\rho_{\text{matter}} < 0.07$.

Further reading: See Chap. 4 in [1] for more details.

Exercises

(1) Calculate the number of photons (N_γ) and entropy in our Universe at recombination and at BBN.

(2) Calculate ρ_γ/ρ_m at BBN and at the electroweak phase transition.
(3) Calculate the ratio $\frac{a_{GUT}}{a_0}$, at the GUT phase transition to today, assuming that standard cosmology follows throughout.
(4) Calculate the number g_* of effectively massless degrees of freedom just below the GUT scale, assuming that only the Standard Model particles are present (and a right-handed neutrino).
(5) Calculate the number density of baryons at recombination.

Chapter 7
The Cosmic Microwave Background Radiation (CMBR) Anisotropy

In this chapter we will describe the Cosmic Microwave Background Radiation and its anisotropy. After describing the CMBR kinematics, we will see how primordial perturbations, in particular curvature perturbations, that were already observed, are described and propagate, and how they influence the CMBR.

7.1 The CMBR and Its Anisotropy

The Cosmic Microwave Background Radiation (CMBR) is a perfect blackbody spectrum, i.e., is distributed according to the Planck distribution,

$$n(\nu)d\nu = \frac{8\pi\nu^2 d\nu}{e^{\frac{h\nu}{k_B T}} - 1}\,, \tag{7.1}$$

present throughout the Universe. As we already mentioned, in the expanding Universe, the Planck distribution remains valid, but the temperature is redshifted like all scales, $T = T(t)$ or rather $T = T(a)$.

The CMBR was discovered in 1965, accidentally, by Arno Penzias and Robert Woodrow Wilson, but it already had a rather long history. It was first postulated by George Gamow in the 1940s, who, as we remember from Chap. 1, had proposed the Hot Universe model, that there should be a thermal blackbody radiation throughout the Universe. He didn't give a definite temperature, but later in the 1940s gave several estimates around 50 K. Then in the late 1940s and early 1950s, Ralph Alpher and Robert Herman, gave an estimate of $T = 5$ K. Robert Dicke estimated it also in the 1950s. The first *published* paper describing it as a *detectable* phenomenon was by A.G. Doroshkevich and Igor Novikov in 1964, but in a Soviet journal, so it was not common knowledge. Jim Peebles, in 1965, had a calculation and a preprint, and based on it, Robert Dicke, Jim Peebles, Roll and David Wilkinson (the one for whom the Wilkinson Microwave Anisotropy Probe, or WMAP, was named) set out

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_7

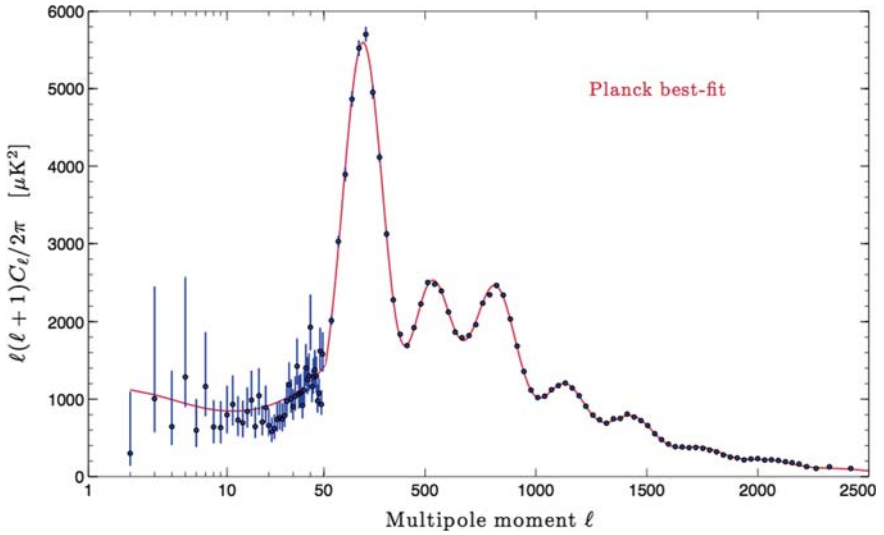

Fig. 7.1 The CMBR spectrum of fluctuations, plotting $l(l+1)C_l/(2\pi)$, quantifying correlations of $\delta T/T$ across an angular scale defined by the angular momentum l, against this l (Image: ESA)

to search for the CMBR, but before they could find it, it was accidentally discovered by Penzias and Wilson, who simply discovered this background radiation that came from extragalactic sources, but had no knowledge of what it means, being unaware of all the history described here. They later learned about the Peebles preprint and the importance of their discovery.

It is found that the CMBR is a blackbody spectrum with amazing accuracy, and the temperature $T(\vec{\Omega})$ seems to be approximately constant as a function of the direction $\vec{\Omega}$ in the sky. However, later on it was found that it is not actually constant, but there is a small *anisotropy* (variation with angle) $\delta T(\vec{\Omega})/T$ of the order of 10^{-5}. The anisotropy is found by the COBE satellite in the 1990s, which is why the absolute value of $\delta T/T$ is still known as the *COBE normalization*. Later, especially with the WMAP and Planck satellites, the CMBR anisotropy was mapped with great precision, and it was found that it exhibits a very rich structure, full of information about the Big Bang.

One plots a quantity $l(l+1)C_l/(2\pi)$, as in Fig. 7.1, roughly describing correlation functions over angular momentum l, against this angular momentum l. There is only one Universe, so the "ensemble" average described here corresponds to the fact that we have divide the sky in many angular distances $\theta \sim 1/l$, and we can correlate them. Indeed, as we will see later, at large l, we can understand l as $\sim 2\pi/\theta$, where θ is the angle. As a result, from simple statistics, the error of the "measurement" $\sim 1/\sqrt{N} \sim 1/\sqrt{l}$, so it is very large at small l. The CMBR plot starts with a plateau, the Sachs-Wolfe plateau (which at small l has huge errors), and has a small l value depending on the Integrated Sachs-Wolfe (ISW) effect. Then we have acoustic peaks, which are damped, so we have a damping tail. This is literally the same as the acoustic damped sinusoidal wave of sound propagation, as we will see later in the chapter.

7.2 Kinematics of the CMBR

The temperature anisotropy is a function of the conformal time η, the spatial position $\vec{x}$ (in our case, the Earth's position in the Universe; if we would have a fast spaceship capable of taking us to a different galaxy, maybe we could vary it, but as it is, we can't), and the angular direction $\vec{n} = \vec{e}$, the direction of the momentum of the photons incoming to us (to the detector), i.e., the angular direction in the sky. We call it

$$\Theta(\eta, \vec{x}, \vec{e}) = \frac{\delta T(\eta, \vec{x}, \vec{e})}{T}. \tag{7.2}$$

But we actually measure intensities of the waves, and because of the Stefan–Boltzmann law, we have $I \propto T^4$, so we actually measure

$$\frac{\delta I}{I} = 4\Theta. \tag{7.3}$$

We expand Θ in multipoles (angular momentum) at each point in spacetime,

$$\Theta(\eta, \vec{x}, \vec{e}) = \sum_{lm} \Theta_{lm}(\eta, \vec{x}) Y_{lm}(\vec{e}). \tag{7.4}$$

Monopole Anisotropy ($l = 0$)

The monopole, i.e., $l = 0$, Θ_{00}, is a local fluctuation (at the position $\vec{x}$ of the Earth), but independent on the direction in the sky. As such, it can't be measured, since all the detectors are at the same position $\vec{x}$ (on the Earth); only if we had a spaceship that could take us to a distant galaxy, we could. This monopole is due to a density perturbation at the current position. Since $\rho_\gamma \propto T^4$, $\Theta = \delta T/T = 1/4\delta\rho_\gamma/\rho_\gamma$, so

$$\Theta_{00}(\eta, \vec{x}) = \frac{1}{4}\frac{\delta\rho_\gamma}{\rho_\gamma}(\eta, \vec{x}). \tag{7.5}$$

Dipole Anisotropy ($l = 1$)

In the CMBR, there is a dipole anisotropy, i.e., for $l = 1$, that is associated with the motion of the observer, i.e. the motion of the Earth, relative to the photon fluid, for CMBR. Note that we are in general relativity, with FRLW expanding metric $ds^2 = -dt^2 + a^2(t)d\vec{x}^2$, instead of the Minkowski metric of special relativity, which means that all frames of reference related by boosts are not equal. There is the special frame of the CMBR, or the *cosmic frame* defined by the time t of the FRLW metric. Indeed, boosting the FRLW metric, we change it, unlike the case of the Minkowski metric.

The terms involving $l = 1$ in the expansion of Θ define the dipole,

$$\sum_m \Theta_{1m} Y_{1m}(\vec{e}) \simeq -\vec{v} \cdot \vec{e} + \cdots = -v \cos\theta + \cdots \,, \tag{7.6}$$

where θ is the angle of the velocity with respect to the angle of observation. Since the energy of the photon gets boosted according to the law

$$E = \gamma(1 + \beta \cos\theta) E' \,, \tag{7.7}$$

it means that the temperature gets changed by

$$T' = \frac{T}{\gamma(1 + \beta \cos\theta)}. \tag{7.8}$$

That gives an expansion in terms of Legendre polynomials,

$$\frac{\delta T}{T} \simeq -\beta P_1(\cos\theta) + \frac{2\beta^2}{3} P_2(\cos\theta) - \frac{\beta^2}{6} + \cdots \tag{7.9}$$

By measuring the dipole, we find the velocity of the motion of the Earth relative to the CMBR,

$$v \simeq 371\,\text{km/s} = 1.2 \times 10^{-3} c. \tag{7.10}$$

But as we saw from the above expansion, besides the dipole (defined by $P_1(\cos\theta)$, we have also contributions at higher $l \geq 2$, with $P_l(\cos\theta)$ induced by the same effect. Just that these contributions are $\mathcal{O}(10^{-6})$, much smaller than the observed 10^{-5} in all l's. That means that the CMBR anisotropy is not just from the dipole, there is really an intrinsic anisotropy of the CMBR.

CMBR Spectrum

We will define the shorthand

$$\Theta_{lm}(\eta_0, \vec{x}_0) \equiv a_{lm}. \tag{7.11}$$

We can consider the "ensemble" average $\langle\rangle$, found by averaging over the sky. We find that

$$\langle a_{lm} \rangle = 0 \,, \tag{7.12}$$

within experimental accuracy, from which we deduce that the Universe is isotropic. Reversely, from rotational invariance (isotropy), we find we must have $\langle a_{lm} \rangle = 0$. Then, under the same assumption of rotational invariance (isotropy), we find that the two-point function depends only on functions C_l,

$$\langle a_{lm} a_{l'm'} \rangle = \delta_{ll'} \delta_{mm'} C_l. \tag{7.13}$$

Substituting in $\Theta(\vec{e})$, we find

$$\langle \Theta(\vec{e}_1)\Theta(\vec{e}_2)\rangle = \sum_l \frac{2l+1}{4\pi} C_l P_l(\cos\theta). \tag{7.14}$$

We can define the variation of C_l,

$$(\Delta C_l)^2 \equiv \langle (|a_{lm}|^2 - C_l)^2\rangle = \langle |a_{lm}|^4\rangle - C_l^2. \tag{7.15}$$

For a *Gaussian* distribution, one obtains

$$(\Delta C_l)^2 = \frac{2C_l^2}{(2l+1)\Delta l}\ , \tag{7.16}$$

since in the numerator the factor of 2 comes from the real and imaginary parts of a_{lm}, and in the denominator $2l+1$ is the number of m's for a_{lm} at given l, and Δl is the binning of l (it is natural to take $\Delta l = 1$, but we can have other choices).

Meaning of Expansion in l and m

The multipole expansion at large l is like a Fourier transform, i.e., is of the type

$$\Theta(\vec{\theta}) = \frac{1}{2\pi}\int d^2\vec{l}\, a(\vec{l}) e^{i\vec{\theta}\cdot\vec{l}}\ , \tag{7.17}$$

so the multipole (angular momentum) number l explores an angle on the sky of

$$\theta \sim \frac{1}{l}. \tag{7.18}$$

The CMBR fluctuations we will be interested in come from given tensor modes. To find the contributions of the various tensor modes, we do a Fourier expansion in the angle of rotation around the direction of the photon momentum $\vec{k}$ (or direction of observation $\vec{e}$). Indeed, the behaviour of a given tensor corresponds to a certain maximum m in the expansion of quantities

$$G(\phi) = \sum_{m\le m_{max}} G_m e^{2\pi i m\phi}. \tag{7.19}$$

Then m_{max} corresponds to the tensor rank: $m_{max} = 0$ is a scalar, $m_{max} = 1$ is a vector, $m_{max} = 2$ is a tensor, etc.

Scalar Mode

But we actually only expect scalar, vector and tensor modes anyway for the perturbations. Moreover, we will focus on the scalar mode in this chapter, since we have a lot of experimental data on it, though there could be also vector and tensor modes, but there is no data yet on this.

Then for the scalar mode, we will only have the $m = 0$ modes, i.e., the $Y_{l0}(\vec{e})$ spherical harmonics. The expansion of Θ for scalar mode contributions only is then, when considering a Fourier expansion over $\vec{x}$ (position), to $\vec{k}$ (momentum)

$$\begin{aligned}\Theta(\eta, \vec{k}, \vec{e}) &= \sum_l (-i)^l \sqrt{4\pi(2l+1)} Y_{l0}(\vec{e}) \Theta_l(\eta, \vec{k}) \\ &= \sum_l (-i)^l (2l+1) P_l[\cos(\hat{k} \cdot \vec{e})]. \end{aligned} \tag{7.20}$$

The reverse of this expansion is

$$\Theta_{lm}(\eta, \vec{x}_0 = 0) \equiv a_{lm} = \frac{4\pi}{(2\pi)^3} i^l \int d^3\vec{k} \Theta(\eta_0, \vec{k}) Y^*_{lm}(\vec{k}) \ , \tag{7.21}$$

and for the scalar mode we have only $\Theta_l = \Theta_{l0} = a_{l0}$. The correlation function in this case is

$$\langle a_{l0} a_{l'0} \rangle = \delta_{ll'} C_l. \tag{7.22}$$

Gaussian Perturbations

Before we proceed, we must describe how to parametrize Gaussian perturbations in some quantity $g(\vec{x})$. Note that of course, non-Gaussian perturbations are important as well (there could be non-Gaussianities in the CMBR), but they have not been discovered yet, so we will ignore them in this chapter.

The variance of g is defined by the correlation function (ensemble average) of g^2,

$$\sigma_g^2(\vec{x}) \equiv \langle g^2(\vec{x}) \rangle \ , \tag{7.23}$$

and it means that the probability to find the value g for $g(\vec{x})$ is

$$P(g) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{g^2}{2\sigma_g^2}}. \tag{7.24}$$

But the variance is expanded in momenta, to find the spectrum of g,

$$\sigma_g^2(\vec{x}) = \frac{1}{(2\pi)^3} \int_0^\infty P_g(k) d^3k = \int_0^\infty \mathcal{P}_g(k) \frac{dk}{k} \ , \tag{7.25}$$

where in the last equality we have considered that if P_g is only a function of the modulus k, we can write a distribution over k,

$$\mathcal{P}_g = \frac{k^3}{2\pi^2} P_g. \tag{7.26}$$

Note that then we have, for the two-point correlation function in *momentum space* (by Fourier transform of the x space result)

$$\langle g_{\vec{k}} g_{\vec{k}'} \rangle = (2\pi)^3 \delta^3_{\vec{k}+\vec{k}'} P_g(\vec{k}). \tag{7.27}$$

Coming back to the case of Θ_l, consider the Gaussian perturbation in $g = \Theta_l = a_{l0}$, for which we have $\sigma_g^2 = \langle (a_{l0})^2 \rangle = C_l$. We obtain therefore for the spectrum

$$C_l = 4\pi \int_0^\infty \mathcal{P}_{\Theta_l}(\eta_0, k) \frac{dk}{k}. \tag{7.28}$$

Moreover, we will be interested in the effect of some primordial perturbation, to be defined shortly, called ζ_k, that is transferred into a perturbation in the CMBR, Θ_l, by the *transfer function* $T_l(k)$,

$$\Theta_l(k) = T_l(k)\zeta_k. \tag{7.29}$$

Replacing in the two-point function,

$$\begin{aligned} \langle \Theta_l^2 \rangle &= \int_0^\infty \mathcal{P}_{\Theta_l}(k) \frac{dk}{k} \\ &= (T_l(k))^2 \langle \zeta_k^2 \rangle = (T_l(k))^2 \int_0^\infty \mathcal{P}_\zeta(k) \frac{dk}{k} , \end{aligned} \tag{7.30}$$

we obtain $\mathcal{P}_{\Theta_l} = (T_l(k))^2 \mathcal{P}_\zeta$, so that

$$C_l = 4\pi \int_0^\infty (T_l(k))^2 \mathcal{P}_\zeta(k) \frac{dk}{k}. \tag{7.31}$$

7.3 Perturbation Theory for CMBR

We now move on to calculating the perturbations in the CMBR. There are in principle several types of perturbations. The one we are interested in, the most important one for its effect on the CMBR, is the *curvature perturbation* ζ, which determines the total density perturbation. There are in principle also isocurvature and tensor perturbations (though they have not been observed yet).

We must consider a gauge for general coordinate transformations, in order to define the perturbations. One considers (fixed) time slices that have uniform density ρ, and threads (fixed x_i worldlines) that are comoving. In that case, we can parametrize the *spatial* metric, perturbed by the curvature perturbation, as

$$g_{ij} = a^2(t) e^{2\zeta(\vec{x},t)} \gamma_{ij}(\vec{x}) , \tag{7.32}$$

where $\det \gamma_{ij} = 1$.

Since these curvature perturbations are determined by the density perturbations, we consider also the density perturbations of the various components,

$$\left(\frac{\delta\rho}{\rho}\right)_{\gamma,\nu,B,c}, \tag{7.33}$$

where γ is for photons, ν for neutrinos, B for baryons and c for Cold Dark Matter.

From the adiabatic condition, the various components depend only on the total density, i.e., the densities are proportional to each other, $\rho_a = \rho_a(\rho)$, and the same is true for the number densities, $n_a = n_a(\rho)$. In particular, that leads to

$$\frac{\delta(n_B/n_\gamma)}{n_B/n_\gamma} = \frac{\delta\rho_B}{\rho_B} - \frac{3}{4}\frac{\delta\rho_\gamma}{\rho_\gamma} = 0. \tag{7.34}$$

More generally, we find also

$$\frac{1}{3}\frac{\delta\rho_B}{\rho_B} = \frac{1}{3}\frac{\delta\rho_c}{\rho_c} = \frac{1}{4}\frac{\delta\rho_\gamma}{\rho_\gamma} = \frac{1}{4}\frac{\delta\rho_\nu}{\rho_\nu}. \tag{7.35}$$

For some generic gauge, transformed from the gauge above (with uniform density slices) to one with $\tilde{t} = t + \delta t(\vec{x}, t)$, we parametrize the spatial metric as

$$g_{ij} = a^2(t)e^{2\psi(\vec{x},t)}\gamma_{ij}(\vec{x}), \tag{7.36}$$

leading to

$$\psi = \zeta - H\delta t = \zeta - H\frac{\delta\rho}{\dot{\rho}} = \psi + \frac{1}{3}\frac{\delta\rho}{\rho + P}, \tag{7.37}$$

where in the last equality we have used the conservation of energy equation.

We next consider perturbations in the fluid velocity.

- Newtonian case. The velocity field is the sum of a Hubble contribution and a peculiar velocity contribution,

$$\vec{u}(\vec{x}, t) = H(t)\vec{r} + \vec{v}(\vec{x}, t). \tag{7.38}$$

  We are mostly interested in a scalar perturbation, which means that the peculiar velocity *in momentum space* satisfies

$$\vec{v}_{\vec{k}}^{\,\text{scalar}} = -i\frac{\vec{k}}{|\vec{k}|}V_{\vec{k}}. \tag{7.39}$$

  Note that the $-i$ is conventional, so that the perturbation makes sense in x space. If we would consider a vector perturbation, that would be defined by $\vec{k}\cdot\vec{v}_{\vec{k}}^{\,\text{vector}} = 0$.
- General relativistic case. The background metric is $ds^2 = a^2(\eta)[-d\eta^2 + \delta_{ij}dx^i dx^j]$, and $dr^i = a\,dx^i$ is the spatial coordinate. Defining the comoving velocity as usual, by

$$u^\mu = \frac{dx^\mu}{d\tau}, \tag{7.40}$$

the normal spatial velocity is

$$v^i = \frac{dr^i}{dt} = a\, u^i. \tag{7.41}$$

Then define the general metric perturbation, independently of (before fixing) any gauge for general coordinate transformations,

$$ds^2 = a^2(\eta)\left\{-(1+2A)d\eta^2 - 2B_i d\eta dx^i + [(1+2D)\delta_{ij} + 2E_{ij}]dx^i dx^j\right\}. \tag{7.42}$$

Here E_{ij} is spatially traceless, since $(1+2D)\delta_{ij}$ is a trace, so this is indeed the most general parametrization.

For a scalar mode, as we are interested in, we would have

$$\begin{aligned} \vec{B} &= -i\frac{\vec{k}}{|\vec{k}|}B \\ E_{ij} &= \left(-\frac{k_i k_j}{k^2} + \frac{1}{3}\delta_{ij}\right)E. \end{aligned} \tag{7.43}$$

However, it is most common to consider the *conformal Newton gauge*, so called since one still has the conformal factor, but except for that, we consider the usual parametrization of small fluctuations in terms of the Newtonian potential, i.e.,

$$ds^2 = a^2(\eta)[-(1+2\Psi)d\eta^2 + (1-2\Phi)\delta_{ij}dx^i dx^j]. \tag{7.44}$$

Note that dropping the conformal factor and replacing $\Psi = \Phi = U_N$ corresponds to the form of the metric depending only on the Newtonian potential U_N, always possible in the small field approximation. Also note that $E = B$ in this gauge.
In this gauge, compared with the uniform density gauge defining ζ, we have from (7.36) expanded to first order that $\psi = -\Phi$, which leads from (7.37) (for $P = w\rho$) to

$$\frac{\delta\rho}{\rho} = 3(1+w)(\zeta + \Phi). \tag{7.45}$$

For a variety of reasons, we know that various scales k must exit the horizon, and spend some time outside the horizon. In inflation, the reason is that the scales are exponentially blown up. During this period, before re-entering the horizon (horizon entry), the perturbations at this scale will be frozen in at some value constant in time. Then, nowadays, during matter domination, the scales come back inside the horizon and then they will start again to evolve in time.
We can find during this period (before horizon entry) that $\delta\rho/\rho = -2\Psi$ (if $\dot{\Phi} = \dot{\Psi} = 0$), which means that

$$-\zeta = \Phi + \frac{2\Psi}{3(1+w)}. \tag{7.46}$$

Time Evolution of Perturbations

In order to find the time evolution of perturbations, we consider as usual the conservation of the energy-momentum tensor $D_\mu T^\mu{}_\nu = 0$, which contains the equations of motion of the fluid.

The 0 component, $D_\mu T^\mu{}_0 = 0$, gives as usual the continuity equation, whereas the i component, $D_\mu T^\mu{}_i = 0$, gives the Euler equation.

Consider a perfect fluid (with no dissipation and no anisotropy), with energy-momentum tensor

$$T^{\mu\nu} = P g^{\mu\nu} + (P+\rho) u^\mu u^\nu. \tag{7.47}$$

Then, the continuity equation for the case of an equation of state $P = w\rho$ for the background gives

$$\left(\frac{\dot{\delta\rho}}{\rho}\right) = -(1+w)(kV - 3\dot{\Phi}) + 3aH\left(w\frac{\delta\rho}{\rho} - \frac{\delta P}{P}\right). \tag{7.48}$$

If we also have $\delta P = w\delta\rho$, which is nontrivial, since it means that fluctuations follow the same equation of state, i.e., that subsystems don't exchange energy or momentum with the exterior (even when there is a local fluctuation), then the last term in the above equation vanishes, and we get

$$\left(\frac{\dot{\delta\rho}}{\rho}\right) = -(1+w)(kV - 3\dot{\Phi}). \tag{7.49}$$

If moreover, also for any component, a subsystem doesn't exchange energy with other subsystem on a given time scale, we have the above relation valid *for any component* as well, i.e.,

$$\begin{aligned}
\left(\frac{\dot{\delta\rho_c}}{\rho_c}\right) &= -kV_c + 3\dot{\Phi}: \quad CDM \\
\left(\frac{\dot{\delta\rho_\gamma}}{\rho_\gamma}\right) &= -\frac{4}{3}kV_\gamma + 4\dot{\Phi}: \quad \text{photon} \\
\left(\frac{\dot{\delta\rho_B}}{\rho_B}\right) &= -kV_B + 3\dot{\Phi}: \quad \text{baryon}
\end{aligned} \tag{7.50}$$

The Euler equation for the case of the perfect fluid becomes now

$$\dot{V} = -aH(1-3w)V - \frac{\dot{w}}{1+w}V + k\frac{\delta P}{\rho + P} + k\Psi. \tag{7.51}$$

It is also true for subsystems (components), in the case that a subsystem doesn't exchange momentum and energy with the exterior (the other subsystems).

Consider the baryon-photon fluid, which is a tightly coupled system until recombination, meaning that the interactions between them are so strong, that they follow the same laws, depending only on the total density. We have then that $V_B = V_\gamma$, and we saw that the perturbation in the baryon energy density is 3/4 of the perturbation in the photon energy density.

Then from the Eq. (7.51) for the baryon and photon components, we can write an equation for the combined fluid, taking into account that $P_\gamma = \rho_\gamma/3$, $P_B = 0$,

$$\dot{V}_\gamma = -aH(1-3\tilde{w})V_\gamma - \frac{\dot{\tilde{w}}}{1+\tilde{w}}V_\gamma + \frac{k}{3}\frac{\rho_\gamma}{\rho_B + \frac{4}{3}\rho_\gamma}\frac{\delta\rho_\gamma}{\rho_\gamma} + k\Psi \ , \tag{7.52}$$

where the equation of state $\tilde{w}$ of the combined baryon-photon fluid is defined as

$$\tilde{w} = \frac{P_{\rm tot}}{\rho_{\rm tot}} = \frac{\rho_\gamma/3}{\rho_\gamma + \rho_B}. \tag{7.53}$$

Eliminating V_γ between (7.51) and (7.52) gives

$$\frac{1}{4}\left(\frac{\ddot{\delta\rho}}{\rho}\right)_{\gamma,k} + \frac{1}{4}\frac{\dot{R}}{1+R}\left(\frac{\dot{\delta\rho}}{\rho}\right)_{\gamma,k} + \frac{k^2}{4}c_s^2\left(\frac{\delta\rho}{\rho}\right)_{\gamma,k} = F_{\vec{k}}(\eta) \ , \tag{7.54}$$

where

$$\begin{aligned} R(\eta) &\equiv \frac{3}{4}\frac{\rho_B}{\rho_\gamma} \\ F_{\vec{k}}(\eta) &\equiv -\frac{k^2}{3}\Psi_k(\eta) + \frac{\dot{R}(\eta)}{1+R(\eta)}\dot{\Phi}_k(\eta) + \ddot{\Phi}_k(\eta) \\ c_s^2(\eta) &\equiv \frac{\partial P}{\partial\rho} = \frac{\dot{P}_\gamma}{\dot{\rho}_\gamma + \dot{\rho}_B} = \frac{1}{3(1+R(\eta))}. \end{aligned} \tag{7.55}$$

Here we see that $R = \delta\rho_B/\delta\rho_\gamma$ and c_s is the speed of sound of the combined fluid, $= dr/d\eta$.

Then the *sound horizon* is

$$r_s(\eta) = \int_0^\eta c_s(\eta)d\eta \ , \tag{7.56}$$

that is, the distance travelled by the sound (in the fluid) from the Big Bang until η. Note that at early times, when there are no baryons yet, $R \simeq 0$, and there is only the radiation, meaning $c_s = 1/\sqrt{3}$, the speed of sound in the photon fluid.

The Eq. (7.54) has the form of damped, forced harmonic oscillator. But in fact, the damping term (with the first derivative of the photon energy density fluctuation)

turns out to be much smaller than the *Silk damping term* to be defined shortly, so we can neglect it.

Then the solution is the solution for the forced harmonic oscillator,

$$\frac{1}{4}\left(\frac{\delta\rho}{\rho}\right)_{\gamma,\vec{k}} = A_{\vec{k}}(\eta) + B_{\vec{k}}(\eta)\cos(kr_s(\eta)) + C_{\vec{k}}(\eta)\sin(kr_s(\eta)). \tag{7.57}$$

Here $A_{\vec{k}}$ is the constant (non-oscillatory) solution of (7.54),

$$A_{\vec{k}}(\eta) = -[1 + R(\eta)]\Phi_k \ , \tag{7.58}$$

as we can explicitly check. These are baryon acoustic oscillations.

In order to find the constant solution however, we need to use a formula that relates the Newtonian potentials Φ and Ψ to the curvature perturbation ζ_k. In the tight coupling approximation for the baryon-photon fluid, we find that $\Phi = \Psi$, which means (from (7.46)) that before horizon entry we find

$$\Phi = \Psi = -\frac{3+3w}{5+3w}\zeta. \tag{7.59}$$

If horizon entry happens close to now, i.e., during matter domination (which is what will happen for most modes observed in the CMBR), when $w = 0$, we find just before horizon entry, i.e., for modes that are on very large scales, that

$$\Phi_k = \Psi_k = -\frac{3}{5}\zeta_k. \tag{7.60}$$

But for modes that have already entered for some time inside the horizon, i.e., for modes that are at smaller scales, or larger l, we have a small modification, parametrized by a *matter transfer function* $T(k)$,

$$\Phi_k = -\frac{3}{5}T(k)\zeta_k. \tag{7.61}$$

Of course, for very large scales $T(k) \simeq 1$.

Silk Damping

The dominant damping effect in the equation for the time evolution of the density perturbation is not the damping term already present, but the so-called Silk damping, which is due to the Thomson scattering of electrons off photons, $\gamma e^- \to \gamma e^-$, which happens at a scale k_D, so gives a factor e^{-k^2/k_D^2} damping, for a solution

$$\frac{1}{4}\left(\frac{\delta\rho}{\rho}\right)_{\gamma,\vec{k}} = A_{\vec{k}}(\eta) + e^{-\frac{k^2}{k_D^2}}[B_{\vec{k}}(\eta)\cos(kr_s(\eta)) + C_{\vec{k}}(\eta)\sin(kr_s(\eta))]. \tag{7.62}$$

One can then solve more precisely the driven, damped harmonic oscillator equation (7.54) and find the coefficients as well, namely

$$\begin{aligned}\frac{1}{4}\left(\frac{\delta\rho}{\rho}\right)_{\gamma,\vec{k}} &= -(1+R)\Phi_k + e^{-\frac{k^2}{k_D^2}}\frac{\zeta_{\vec{k}}}{3}\cos(kr_s(\eta)) \Rightarrow \\ \frac{1}{4}\dot{\left(\frac{\delta\rho}{\rho}\right)}_{\gamma,\vec{k}} &= -kc_s e^{-\frac{k^2}{k_D^2}}\frac{\zeta_{\vec{k}}}{3}\sin(kr_s(\eta)) \ ,\end{aligned} \tag{7.63}$$

where $\Phi_k = -3T(k)\zeta_k/5$.

7.4 Effect of Density Perturbations on the CMBR

We now go back to the CMBR, to find the effect of the density perturbation above on the CMBR. Consider the *sudden decoupling approximation* for the CMBR. This means that at recombination, decoupling happens on a given time η_{ls} (ls stands for last scattering, of photons off electrons) for all the photons at the same time. Then *at last scattering*, the only anisotropy we have is due to the monopole (local density perturbation at the position we consider) plus dipole, i.e.,

$$\Theta_{ls}(\vec{e}) = \left(\frac{1}{4}\frac{\delta\rho_\gamma}{\rho_\gamma} + \vec{e}\cdot\vec{v}_\gamma\right)_{ls}. \tag{7.64}$$

Sachs-Wolfe Effect

The dependence on direction happens due to the photons travelling since last scattering until now, through the so-called Sachs-Wolfe effect, so

$$\Theta(\vec{e}) = \Theta_{ls}(\vec{e}) + \Theta_{SW}(\vec{e}). \tag{7.65}$$

Consider a 4-momentum (p^0, p^i) in a locally orthogonal (Minkowski) frame, described as (q, qu^i). On the other hand, in the conformal Newton gauge, the 4-momentum is

$$p^0 = (1+\Psi)q; \quad p^i = (1-\Phi)qu^i. \tag{7.66}$$

Then

$$0 = -(p^0)^2 + p^i p^i \simeq -(1+2\Psi)q^2 + (1-2\Phi)q^2u^iu^i = q^\mu q^\nu g_{\mu\nu}. \tag{7.67}$$

The Sachs-Wolfe contribution comes from the redshift of nearby photons. The momentum field $q = q(\vec{x}, t)$ is written, by taking out the expanding factor, as

$$q(\vec{x}, t) = a(t)p(\vec{x}, t). \tag{7.68}$$

The redshift in the comoving frame p has therefore two contributions,

$$\frac{dp}{p} = -\frac{da}{a} + \frac{dq}{q} \ , \tag{7.69}$$

where the first term is the Hubble term (coming from the expansion of the Universe), and the second is the effect we are looking for, coming from the change in the metric of spacetime due to the perturbations,

$$\Theta_{SW} = \int_{\eta_{ls}}^{\eta_0} \frac{dq}{q} = \int_{\eta_{ls}}^{\eta_0} d\eta \frac{1}{q}\frac{dq}{d\eta}. \tag{7.70}$$

To find this expression, we consider the geodesic equation in curved space for the motion of the photon,

$$\frac{dq^\mu}{dx^0} = g^{\mu\nu}\left(\frac{1}{2}\partial_\nu g_{\alpha\beta} - \partial_\beta g_{\nu\alpha}\right)\frac{q^\alpha q^\beta}{q^0}. \tag{7.71}$$

Considering the 0 component for perturbations, and dividing by $q \equiv q^0$, and writing it in terms of η, we obtain

$$\frac{1}{q}\frac{dq}{d\eta} = \frac{\partial\Phi}{\partial\eta} - u^i\frac{d\Psi}{dx^i} \ ., \tag{7.72}$$

where $u^i = dx^i/d\eta$. Since

$$\frac{d\Psi}{d\eta} = \frac{\partial\Psi}{\partial\eta} + \frac{dx^i}{d\eta}\frac{\partial\Psi}{\partial x^i} \ , \tag{7.73}$$

we obtain

$$\frac{1}{q}\frac{dq}{d\eta} = \frac{\partial(\Phi+\Psi)}{\partial\eta} - \frac{d\Psi}{d\eta}. \tag{7.74}$$

Then the Sachs-Wolfe contribution is

$$\Theta_{SW}(\vec{e}) = \Psi_{ls} - \Psi_0 + \int_{\eta_{ls}}^{\eta_0} d\eta \frac{\partial}{\partial\eta}(\Phi+\Psi). \tag{7.75}$$

In this formula, the $-\Psi_0$ term is usually neglected, since nowadays the fluctuation on very large scales in the Universe is negligible with respect to Ψ_{ls}. Then Ψ_{ls} is a contribution from the surface of last scattering, like the non-SW contribution of $1/4\delta\rho_\gamma/\rho_\gamma$, and the last contribution is an *integrated Sachs-Wolfe* (ISW) contribution.

Sachs-Wolfe Plateau

At small l, i.e., for the largest scales, for which the mode has barely left the horizon, during the matter dominated epoch, the mode has spent most of the time since

decoupling to present outside the horizon, where it is constant, i.e., η-independent, i.e., no ISW contribution. Since moreover during this time $\Psi_{ls} = \Phi_{ls} = \Theta_{SW}(\vec{e})$, we find

$$\Theta(\vec{e}) = \frac{1}{4}\left(\frac{\delta\rho_\gamma}{\rho_\gamma}\right)_{ls} + \Phi_{ls}. \tag{7.76}$$

In this case, one finds that $l(l+1)C_l$, the quantity plotted in the CMBR spectrum, is approximately independent of l (flat), leading to the SW plateau, until about $l \sim 30$. Since we find $\delta\rho_\gamma/\rho_\gamma = -8\Phi/3$, we find

$$\Theta(\vec{e}) = \frac{1}{3}\Phi_{ls} = \frac{1}{5}\zeta_{ls}. \tag{7.77}$$

Integrated SW Effect

The integral of the ISW has contributions coming from radiation still present at decoupling, as well as, during late times, from the cosmological constant (dark energy).

Acoustic Peaks and Silk Damping

For l greater than about 30, we have a snapshot of acoustic oscillations at decoupling.

Substituting inside the definition

$$\Theta_l(\vec{k}) = -i^l[4\pi(2l+1)]^{-1/2}\int Y_{l0}^*(\vec{e})\Theta(\eta, \vec{k}, \vec{e})d^2\vec{e} \tag{7.78}$$

the formula for Θ, we find

$$\Theta_l(\vec{k}) = \left[\frac{1}{4}\left(\frac{\delta\rho_\gamma}{\rho_\gamma}\right)_{\vec{k}} + \Phi_{\vec{k}}\right]_{ls} j_l(k\eta_0) \equiv T_l(k)\zeta_k\ , \tag{7.79}$$

where j_l are the Bessel functions. Substituting (7.63), we find

$$T_l(k) = \left[\frac{3}{5}RT(k) - \frac{1}{3}e^{-\frac{k^2}{k_D^2}}\cos kr_s\right] j_l(k\eta_0). \tag{7.80}$$

But in the formula

$$C_l = 4\pi\int_0^\infty T_l^2(k)\mathcal{P}_\zeta(k)\frac{dk}{k}\ , \tag{7.81}$$

one finds (observationally) that the spectrum $\mathcal{P}_\zeta(k)$ is approximately flat (constant), so that

$$C_l = 4\pi\mathcal{P}_\zeta\int_0^\infty\left[\frac{3}{5}RT(k) - \frac{1}{3}e^{-\frac{k^2}{k_D^2}}\cos kr_s\right]^2 j_l^2(k\eta_0)\frac{dk}{k}. \tag{7.82}$$

This shows the presence of acoustic oscillations and Silk damping. Here, from observations plus theory, we find that

$$r_s \simeq 150\,\text{Mpc}; \quad \eta_0 \simeq 14000\,\text{Mpc}; \quad k_D^{-1} \simeq 8\,\text{Mpc}. \tag{7.83}$$

Note that in this chapter we haven't described polarization modes, isocurvature or tensor modes, which could nevertheless become important, if they are observed.

Important Concepts to Remember

- The CMBR, the radiation coming from the Big Bang from every point in the Universe, is approximately homogenous and isotropic, but it has anisotropic fluctuations of $\delta T/T \sim 10^{-5}$.
- The CMBR has a monopole anisotropy at our current position in the Universe, but that can't be measured.
- The CMBR has also a dipole anisotropy due to the Earth's motion with respect to the cosmological frame (frame of the CMBR).
- The CMBR spectrum is found as an expansion over angles, leading to an expansion in multipoles l. At large l, $l \sim 1/\theta$.
- Gaussian perturbations have 2-point correlator $\langle g_{\vec{k}} g_{\vec{k}'} \rangle = (2\pi)^3 P_g(\vec{k})\delta(\vec{k} + \vec{k}')$ and $\mathcal{P}_g(k) = k^3/(2\pi^2)P_g(\vec{k})$.
- There can be scalar, vector and tensor perturbations, but we focus on the scalar ones.
- Primordial perturbations ζ_k are transferred to the CMBR as $\Theta_l(k) = T_l(k)\zeta_k$. Here ζ_k is the curvature perturbation, related to the density perturbations.
- In the conformal Newton gauge, the perturbations are defined by the Newtonian potential, $\Psi = \Phi = U_N$.
- We are interested in perturbations that have exited outside the horizon early on, so that they can come back inside during current times, as CMBR perturbations.
- We find that the density perturbation created by ζ is $1/4(\delta\rho/\rho)_{\zeta,k} = A_k(\eta) + e^{-k^2/k_D^2}[B_k(\eta)\cos(kr_s(\eta)) + C_k(\eta)\sin(kr_s(\eta))]$, giving oscillations and Silk damping.
- In the sudden decoupling approximation, the above perturbation is translated by the propagation of light into the CMBR perturbation, and one finds that an approximately flat $\mathcal{P}_\zeta(k)$ (spectrum of curvature perturbations) generates the damped oscillations of the CMBR.

Further reading: See Chaps. 5, 6, 7, 8 and 10 in [9].

Exercises

(1) Prove that for $P = w\rho$ (for the background), the continuity equation gives

$$\left(\frac{\dot{\delta\rho}}{\rho}\right) = -(1+w)(kV - 3\dot{\Phi}) + 3aHw\left(\frac{\delta\rho}{\rho} - \frac{\delta P}{P}\right). \tag{7.84}$$

(2) Prove that the Euler equation gives

$$\dot{V} = -aH(1-3w)V - \frac{\dot{w}}{1+w}V + k\frac{\delta P}{\rho + P} + \kappa\Psi. \tag{7.85}$$

(3) Eliminate V_γ between

$$\left(\frac{\dot{\delta\rho_\gamma}}{\rho_\gamma}\right) = -\frac{4}{3}kV_\gamma + 4\dot{\Phi} \tag{7.86}$$

and

$$\dot{V}_\gamma = -aH(1-3\tilde{w})V_\gamma - \frac{\dot{\tilde{w}}}{1+\tilde{w}}V_\gamma + \frac{k}{3}\frac{\rho_\gamma}{\rho_B + \frac{4}{3}\rho_\gamma}\frac{\delta\rho_\gamma}{\rho_\gamma} + k\Psi\ , \tag{7.87}$$

to find

$$\frac{1}{4}\left(\frac{\ddot{\delta\rho}}{\rho}\right)_{\gamma,k} + \frac{1}{4}\frac{\dot{R}}{1+R}\left(\frac{\dot{\delta\rho}}{\rho}\right)_{\gamma,k} + \frac{k^2}{4}c_s^2\left(\frac{\delta\rho}{\rho}\right)_{\gamma,k} = F_{\vec{k}}(\eta). \tag{7.88}$$

(4) Calculate the time evolution of the speed of sound c_s from recombination until today, and then of the sound horizon $r_s(\eta)$.

Chapter 8
Problems to Be Solved by Inflation and How They Are Solved in Inflationary Models

In this chapter we will present problems with the standard cosmology, before the advent of the inflationary paradigm, and then we will see how they are solved by inflation.

8.1 Problems with Standard Cosmology Before Inflation

1. Smoothness and horizon problem

Why is the Universe uniform and isotropic?

This problem appeared because the light that we observe in the CMBR, coming from the surface of last scattering (recombination), when considered as coming from *opposite* directions on the sky, comes from regions of the Universe that are causally disconnected, since the horizon distance is smaller than the distance travelled by light.

That sounds strange, since causality is defined also by light propagation, but it is connected with the fact that the Universe expands while light propagates, and the fact that causal contact refers *to the moment of last scattering* (so light propagation before it), whereas light propagation for the CMBR refers to the time after last scattering.

The fact that the light comes from seemingly causally disconnected regions would imply that there would be no reason for the correlation implied by the uniform and isotropic temperature and Planck distribution. Since T is constant to a high degree of precision, and moreover, there are correlations in the sky even in the fluctuations, it seems to say that the surface of last scattering in opposite directions with respect to us was causally connected. We will see that this is what happens in inflation.

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_8

We now turn to a quantitative evaluation of what we need to fix. The horizon distance at the time of last scattering t_{ls} is defined by Δx for light propagation from $t = 0$ (Big Bang) and t_{ls},

$$d_H(t_{ls}) = a(t_{ls})\Delta x = a(t_{ls}) \int_0^{t_{ls}} \frac{dt}{a(t)}. \tag{8.1}$$

This distance, translated into today's distances, by replacing $a(t_{ls})$ with $a(t_0)$, is

$$d_H(t_0) = a(t_0) \int_0^{t_{ls}} \frac{dt}{a(t)}. \tag{8.2}$$

On the other hand, the distance travelled by light from the surface of last scattering until today, in today's scales, is

$$r_H(t_0) = a(t_0) \int_{t_{ls}}^{t_0} \frac{dt}{a(t)}. \tag{8.3}$$

Light propagates during this time in a matter dominated Universe, when $a \propto t^{2/3}$, or more precisely $a(t) = a(t_0)(t/t_0)^{2/3}$, so that

$$r_H(t_0) = 3(t_0^{1/3} - t_{ls}^{1/3})t_0^{2/3}. \tag{8.4}$$

Even though on a log scale, matter domination takes a very small piece of the Universe before last scattering, in absolute terms it dominates (as we saw in Chap. 6, $t_{ls} \simeq 200 t_{eq}$), so we can also use M.D. to compute $d_H(t_{ls})$, giving

$$d_H(t_0) = 3t_{ls}^{1/3} t_0^{2/3}\ , \tag{8.5}$$

so finally the ratio we want, of the distance between the opposite last scattering points vs. the horizon distance, is

$$N = \frac{2r_H(t_0)}{d_H(t_0)} \simeq 2\frac{t_0^{1/3} - t_{ls}^{1/3}}{t_{ls}^{1/3}} \simeq 2\left(\frac{t_0}{t_{ls}}\right)^{1/3} = \left(\frac{a_0}{a_{ls}}\right)^{1/2} = (1 + z_{ls})^{1/2}. \tag{8.6}$$

Here we have used the fact that $t_0 \gg t_{ls}$, that $a(t) \propto t^{2/3}$ during matter domination, and that $a \propto 1 + z$, where z is the redshift. Putting in numbers, specifically that $z_{ls} \simeq 1300$, as we calculated in Chap. 6, we find

$$N = 2 \times \sqrt{1300} \simeq 72. \tag{8.7}$$

So we need to find a way to increase the horizon distance at the surface of last scattering by a factor of 72 with respect to the naive result.

2. Flatness problem

Why do we have $\Omega \simeq 1$ *in the past?*

We have seen that the Friedmann equation can be written in as

$$\Omega(t) - 1 = \frac{K}{a^2(t)H^2(t)} = \frac{K}{a^2(t)} \frac{3\Omega}{8\pi G\rho(t)} , \tag{8.8}$$

where in the last equality we have used $H^2 = 8\pi G\rho_{\rm cr}/3$ and $\Omega = \rho/\rho_{\rm cr}$. Then we obtain

$$\frac{\Omega(t) - 1}{\Omega(t)} = \frac{3K}{8\pi G\rho(t)a^2(t)}. \tag{8.9}$$

Since $\rho(t) \propto 1/t^2$ and $a(t) \propto t^p$, where $p = 2/3$ for M.D. and $p = 1/2$ for R.D., we find that

$$\frac{\Omega(t) - 1}{\Omega(t)} \propto t^{2(1-p)}. \tag{8.10}$$

In turn that means that even if $\Omega \neq 1$, so $K \neq 0$ (K is ± 1), considering an Ω which is of order 1, but not 1, nowadays, it will be much closer to 1 in the past, to an absurd precision. For instance, $\Omega \sim 1$ now implies $\Omega - 1 \sim 10^{-16}$ at e^+e^- annihilation, and even smaller at earlier times.

But that begs a question: why is it so unbelievably small then? Since what it is now, is a result of what it was in the past, we need to explain this unnaturally small number (saying it is what it is, via the most vanilla type of anthropic principle, the fact that the Universe is as it is, since otherwise there would be no people to see it, simply avoids any attempt at an explanation).

Of course, it could be that we have exactly $\Omega \equiv 1$, in which case $K = 0$, but we have no *experimental* evidence for that. Saying that $\Omega = 1$ would be a theoretical bias; experimentally we only know that $\Omega \simeq 1$ to about one percent, so we must assume the equality is not perfect.

But then we need to explain the absurdly small number for $\Omega - 1$ at very early times.

3. Why is the entropy so large?

We have already seen that the entropy per baryon is about $S_1 \sim 10^9$. Multiplying with the number of baryons within the current horizon, or alternatively calculating the entropy density s and multiplying by the horizon distance volume, $S = s \cdot \frac{4\pi}{3}(ct)^3$, we find the entropy $S \sim 10^{88}$ today.

But actually we need ot consider the entropy at some earlier time, when it is actually smaller, since $s \propto T^4$, but $(ct)^3 \propto (1/T^2)^3$ during the R.D. era and $(ct)^3 \propto (1/T^{3/2})^3$ during the M.D. era. For instance, at BBN, one finds $S_H(t_{BBN}) \sim 10^{63}$, which is still huge.

We need to explain the great production of entropy, since naturally we would expect again numbers of order one for the entropy within the horizon, or the entropy per baryon ratio.

4. Why do we have a (very small) baryon asymmetry?

As we have seen, we have an asymmetry

$$\frac{N_B - N_{\bar{B}}}{N_B} \sim \frac{1}{10^9} , \tag{8.11}$$

which needs to be explained (why it is nonzero, and why it is so small).

5. How do we generate perturbations?

This is sort of the opposite of the question at number one, in the sense that we see an Universe which is not just approximately uniform and isotropic, but also full of small fluctuations. On the largest scales, we have the fluctuations in the CMBR; on smaller scales, the fluctuations that have since then evolved to give rise to structure, namely stars and galaxies. Note that these seem to be *classical* fluctuations, and moreover on scales that were *super-horizon* in the past, i.e., they should not have been in causal contact. Indeed, scales grow as $a(t)$, but the horizon grows like ct, which during R.D. is $a(t)^2$, and during M.D. is $a(t)^{3/2}$, growing faster than the scales. Thus scales keep falling inside the horizon as time evolves: scales that were super-horizon drop to sub-horizon, where we can observe them. This is most puzzling in standard cosmology, and seems to have no explanation.

6. Monopole problem

Why so small relic densities in the Universe?

Historically, the problem referred to monopoles, which are creating during a phase transition, like a GUT phase transition, so we expect to find them in large numbers. Nowadays, we say that any kind of relics, like gravitinos, string moduli, or other topological defects (e.g., cosmic strings, domain walls, etc.), would be produced in large numbers at early times. But since we don't observe them in large numbers, there must be a mechanism to dilute them.

By the *Kibble mechanism*, we expect to create one monopole per horizon during a phase transition, as the Higgs VEV obtains a random but uniform value in a region of the order of the horizon, as the Universe cools through the phase transition, and a random set of scalar VEVs on horizons would generically create a monopole configuration for these scalars. But during the same GUT phase transition, we also expect to create one nucleon per horizon by similar arguments. Since neither decays much until now, we would expect to see about one monopole per nucleon, or one per 10^9 photons. However, direct monopole searches in materials on Earth show that there are less than 10^{-30} monopoles per nucleon, which means we should reduce the monopole density by about 10^{-30} in volume. For general relics (like for instance cosmic strings), one can calculate that in order for their mass density not to over-close the Universe, we need a reduction by a factor of about 10^{-11}, which is less stringent than the monopoles' 10^{-30}, but still important.

8.2 The Paradigm of Inflation

The paradigm that comes to solve all of the above problems is called *inflation*. It is more of a paradigm than a theory, since it is a set of very general assumptions. It was started in the West by Alan Guth in 1981, with the model now known as "old inflation", which however had some problems, to be described later. It was then refined to what is now known as "new inflation" in the independent works by Albrecht and Paul Steinhardt in 1982, and by Andrei Linde in 1982 and 1983. However, it is worth pointing out that the model that fits the current data best is the model that came before Guth, of "Starobinsky inflation", a modification of general relativity now known to be equivalent with a scalar model with a certain exponential potential. Starobinsky wrote his model in 1979 in Soviet Russia (in JETP Letters), but it didn't have an impact in the West, and was only later re-discovered.

The paradigm can be stated most generally as the need to have a (long) period of accelerated inflation, $\ddot{a} > 0$. Since the acceleration equation is

$$\frac{\ddot{a}}{a} = -\frac{4\pi G_N}{3}(\rho + P) \ , \tag{8.12}$$

we need $\rho + P < 0$ or, since $P = w\rho$, that $w < -1/3$. The most common way is to have also an *exponential* expansion, through an approximately constant energy density, but we don't actually need that in order to solve the problems. It is also common to assume that the approximately constant energy density is parametrized by a scalar field, but this is only the simplest case, not a necessity. In principle, we could have any sort of tensor field, e.g. vector, graviton, higher spin, that parametrizes the change in the energy density.

Then the standard incarnation of inflation involves an energy density that is approximately constant, so a dark energy component that is approximated by a cosmological constant Λ. We will review the energetics of Λ, that we already touched on before. The gravitational acceleration felt at a radius R is, from the acceleration equation,

$$\ddot{R} \equiv g_{\rm GR} = -\frac{4\pi G}{3} R \left(\rho + \frac{3P}{c^2}\right) \ , \tag{8.13}$$

whereas in the Newtonian case we have the same formula, but with no P contribution,

$$\ddot{R} \equiv g_{\rm Newton} = -\frac{4\pi G}{3} R\rho. \tag{8.14}$$

For a cosmological constant, $P_V/c^2 = -\rho_V$, which leads to the positive gravitational acceleration

$$g_V = +\frac{8\pi G}{3} R\rho_V \ , \tag{8.15}$$

which means a *repulsive gravity*. The reason that is possible is because the negative pressure creates positive work. Indeed, we have (for $c = 1$) $dE = -PdV$, but $E = \rho V$, which means that if $P = -\rho$, then $d\rho = 0$, so the energy density is constant

$$\rho_V = \frac{\Lambda}{8\pi G} = \text{const.} \tag{8.16}$$

as the volume expands. That means that energy is being created, by the work of the negative pressure (normally, as the volume expands, energy is lost, or at most constant, through the work of the positive pressure).

From the acceleration equation,

$$\ddot{a} = \frac{8\pi G\rho_V}{3} a \ , \tag{8.17}$$

whereas the Friedmann equation gives

$$H^2 = \frac{8\pi G}{3}\rho_{\text{tot}} \ , \tag{8.18}$$

leading to an exponential expansion,

$$a(t) = a_0 e^{H_0 t} \ , \tag{8.19}$$

where the *constant* Hubble scale H_0 is

$$H_0 = \sqrt{\frac{8\pi G\rho_V}{3}} = \sqrt{\frac{\Lambda}{3}}. \tag{8.20}$$

To finalize the description, we need to state that the exponential expansion happens near the Planck scale. The Planck scale is defined as the combination of c, $\hbar$ and G that has the right dimensions, so the Planck time and length are

$$t_P = \sqrt{\frac{\hbar G}{c^3}} \sim 10^{-43}\,\text{s}, \quad l_P = ct_P \sim 10^{-35}\,\text{m} \ , \tag{8.21}$$

whereas the Planck mass is

$$m_P = \sqrt{\frac{c\hbar}{G}} = 10^{19}\,\text{GeV} = 10^{-8}\,\text{Kg}. \tag{8.22}$$

Note however that the Planck mass m_P above is the quantity defined in particle physics, but in gravity and cosmology one uses instead a different definition, the *reduced Planck mass* M_{Pl}, defined as (for $c = \hbar = 1$)

$$M_{\rm Pl} = \frac{1}{\sqrt{8\pi G}} = \frac{m_P}{\sqrt{8\pi}}. \tag{8.23}$$

The definition is taken such that the coefficient of the Einstein–Hilbert action $\int d^4x\sqrt{-g}R$ is $\frac{M_{\rm Pl}^2}{2}$, meaning that the graviton modes have the standard canonical scalar kinetic action.

As we said, only accelerated expansion is enough for the inflation paradigm, which means that we could also have the so-called *power-law inflation*, with

$$a(t) \propto t^p, \tag{8.24}$$

where $p > 1$. Indeed, in that case, we obtain $\ddot{a} > 0$.

8.3 Inflationary Solutions to Cosmological Problems

We now turn to how to solve the problems mentioned with standard cosmology within inflation.

Flatness Problem

We saw that

$$\Omega(t) - 1 = \frac{K}{(aH)^2} \sim \left(\frac{t}{a}\right)^2 \propto t^{2(1-p)}. \tag{8.25}$$

In the case of standard cosmology, with $p < 1$ (matter domination has $p = 2/3$, radiation domination has $p = 1/2$), $\Omega(t) - 1$ grows with time. But by having a long period of accelerated expansion, i.e., inflation, with $p > 1$, we could fix it.

For instance, if we consider the case of cosmological constant (considered by Guth), with $H = H_I$ constant and $a(t) \propto e^{Ht}$,

$$\Omega(t) - 1 = \frac{K}{a^2H^2} \propto \frac{K}{a^2} \propto e^{-2H_I t}\ , \tag{8.26}$$

which drops exponentially fast. That means that by starting with an $\Omega - 1 \sim \mathcal{O}(1)$, it drops to a very small value after some e-folds, so it can then grow again during the radiation domination and matter domination eras. Though typically, in this way we still obtain today a very small $\Omega - 1$, effectively zero. This is so since it would be very strange that after coming to an exponentially small value, it would grow back to a value of order one just now.

So the theoretical bias that we talked about is based on inflation, when it is natural to obtain an exponentially small $\Omega(t_0) - 1$ (nowadays). That is why we believe that Ω is one to an excellent accuracy now, even though experiments only show it to a small accuracy.

We now turn to a quantitative analysis, within exponential inflation, that will aim to constrain the amount of inflation needed. To do so, we we define the *number of e-foldings* N_e during inflation,

$$\frac{a(t_I)}{a(t_{bi})} = e^{2H_I t} \equiv e^{N_e} \rightarrow N_e = \ln \frac{a(t_I)}{a(t_{bi})} , \tag{8.27}$$

where t_{bi} is the the of begining of inflation, and t_I is the time of end of inflation.

Then now (at time t_0), we have

$$\Omega_0 - 1 = \frac{K}{a_0^2 H_0^2} = \frac{K}{a_I H_I^2} = \frac{K}{a_{bi}^2 H_{bi}^2} e^{-2N_e} \left(\frac{a_I H_I}{a_0 H_0} \right)^2 , \tag{8.28}$$

and if we assume $K/(a_{bi} H_{bi})^2 = \Omega(t_{bi}) - 1 \sim \mathcal{O}(1)$, then in order to solve the flatness problem, we need

$$e^{N_e} > \frac{a_I H_I}{a_0 H_0}. \tag{8.29}$$

But we consider that the time of begining of radiation domination, denoted by index 1, is just after the end of inflation, so that $a_I H_I \simeq a_1 H_1$. In reality it could be a couple of e-folds later, but it doesn't matter much for the calculation.

Since during radiation domination $H \propto 1/t \propto 1/a^2$, and the end of radiation domination corresponds to the matter-radiation equality, when $H^2 \propto \rho = \rho_m + \rho_{\rm rad} = 2\rho_{\rm rad} \propto 2H_{rad}'^2$, we obtain

$$H_1 = H' \left(\frac{a_{eq}}{a_1} \right)^2 = \frac{H_{eq}}{\sqrt{2}} \left(\frac{a_{eq}}{a_1} \right)^2 . \tag{8.30}$$

On the other hand, during radiation domination, $H \propto 1/t \propto 1/a^{3/2}$, and the end of matter domination corresponds to nowadays, when $H^2 \propto \rho_{0,\rm cr} = \rho_{m,0}/\Omega_{m,0}$, where $\Omega_{m,0} = \rho_{m,0}/\rho_{0,\rm cr}$, we similarly obtain

$$H_{eq} = \sqrt{2\Omega_{m,0}} H_0 \left(\frac{a_0}{a_{eq}} \right)^{3/2} . \tag{8.31}$$

Then the condition on the number of e-folds is

$$e^{N_e} > \frac{a_I H_I}{a_0 H_0} = \frac{a_1 H_1}{a_0 H_0} = \frac{a_1 \sqrt{H_1}}{a_0 \sqrt{H_0}} \sqrt{\frac{H_1}{H_0}} , \tag{8.32}$$

and since $a_1 \sqrt{H_1} = 2^{-1/4} a_{eq} \sqrt{H_{eq}}$, and $a_{eq} \sqrt{H_{eq}} = a_0 \sqrt{H_0} \sqrt{2\Omega_{m,0}} \sqrt{a_0/a_{eq}}$, we obtain

$$e^{N_e} > \left(\Omega_{m,0} \frac{a_{eq}}{a_0} \right)^{1/4} \sqrt{\frac{H_1}{H_0}} = \Omega_{\rm rad,0}^{1/4} \sqrt{\frac{H_1}{H_0}} = \left(\Omega_{\rm rad,0} \frac{\rho_1}{\rho_{0,\rm cr}} \right)^{1/4} , \tag{8.33}$$

where we used $H^2 = (8\pi G/3)\rho$ if $K/(a^2H^2) \to 0$ (like in the case of ρ_1, the density at the begining of the RD era), or $H^2 = (8\pi G/3)\rho_{0,\rm cr}$. We have also used the fact that $\Omega_m/\Omega_{\rm rad} = \rho_m/\rho_{\rm rad} \propto a$, so $\Omega_{m,0}/\Omega_{\rm rad,0} = \Omega_{\rm m,eq}/\Omega_{\rm rad,eq} a_0/a_{eq} = a_0/a_{eq}$.

We then substitute numbers, giving

$$\rho_{\rm cr,0} = [3 \times 10^{-3}\,{\rm eV}]^4 h^2; \quad \text{and} \quad \Omega_{\rm rad,0} \simeq \frac{\Omega_{m,0}}{20000} \sim \frac{0.3}{20000}\,, \tag{8.34}$$

so that the bound is ($h \sim 0.7$)

$$e^{N_e} > \frac{[\rho_1]^{1/4}}{0.037\,{\rm heV}} \sim e^{56} \frac{[\rho_1]^{1/4}}{5 \times 10^{13}\,{\rm GeV}}. \tag{8.35}$$

On the other hand, the scale $[\rho_1]^{1/4}$, approximately the scale of inflation, is bounded from below by the Big Bang Nucleosynthesis scale (since we know everything that happens from BBN onwards, and inflation has no place in it), so $[\rho_1]^{1/4} > \rho_{BBN}^{1/4} \sim 1\,{\rm MeV}$. For this minimal value, the bound would be

$$e^{N_e} > e^{56} \times 2 \times 10^{-17} \simeq e^{17}. \tag{8.36}$$

The scale is also bounded from above by the Planck scale, $[\rho_1]^{1/4} < [\rho_{\rm Pl}]^{1/4} \sim 10^{19}\,{\rm GeV}$. For this maximal value, the bound would be

$$e^{N_e} > e^{56} \times 2 \times 10^{7} \simeq e^{68}. \tag{8.37}$$

In fact, the value preferred by experimental constraints on generic models is $[\rho_1]^{1/4} \sim 2 \times 10^{16}\,{\rm GeV}$, so that the bound in this case is

$$e^{N_e} > e^{56} \times 10^{3} \simeq e^{62}. \tag{8.38}$$

Later on in the book, we will see a more precise formula, which is

$$N_e > 56 - \frac{2}{3} \ln \frac{10^{16}\,{\rm GeV}}{\rho_*^{1/4}} - \frac{1}{3} \ln \frac{10^{19}{\rm GeV}}{T_R}\,, \tag{8.39}$$

where T_R is the reheat temperature, and ρ_* is the density at the end of inflation. Then for the case $T_R = \rho_*^{1/4}$, we obtain

$$N_e > 56 - \ln \frac{10^{13.6}{\rm GeV}}{\rho_*^{1/4}} = 56 + \ln \frac{\rho_*^{1/4}}{5 \times 10^{13}{\rm GeV}}\,, \tag{8.40}$$

which is the previous constraint for $\rho_* \to \rho_1$.

Smoothness and Horizon Problem

We have seen that during the radiation dominated era, with horizon distance $H^{-1} \sim t \propto a^2$, and during the matter dominated era, with horizon distance $H^{-1} \sim t \propto a^{3/2}$, scales enter within the horizon, since scales go like a.

The solution to the smoothness and horizon problem is that during the period of exponential expansion with $H^{-1} =$ constant (or even just accelerated expansion $a(t) \sim t^p$, $p > 1$, with $H^{-1} \propto t \propto a^{1/p}$), scales $\sim a$ get outside the horizon.

So the picture is that a small patch of space is blown up exponentially by the expansion, while the horizon distance remains constant, and in so doing, the scale gets "frozen-in", losing causal contact. Then during either the R.D. or M.D. eras, the scales falls back within the horizon, and regains causal contact, starting to react and evolve again. Then the fact that there was an apparent lack of causal contact on the largest scales is simply because the scales were frozen-in, but they actually used to be in causal contact before getting outside the horizon.

To describe the horizon problem in quantitative terms, we define first, as before, the horizon distance at time t_{ls},

$$d_H(t_{ls}) = a(t_{ls}) \int_{t_{bi}}^{t_{ls}} \frac{dt}{a(t)} \,, \tag{8.41}$$

which is dominated by the period of inflation begining at t_{bi}. That sounds counterintuitive, since we just argued that simply a factor of 200 in time during matter domination means that the latter period dominates the integral, but we will see that the magic of exponential expansion means that the early times actually dominate the integral. During inflation, for $t < t_I$,

$$a(t) = a_I e^{-H_I(t_I - t)} \,, \tag{8.42}$$

and the number of e-folds during inflation is

$$N_e = H_I(t_I - t_{bi}) \,, \tag{8.43}$$

leading to

$$d_H(t_{ls}) \simeq \frac{a(t_{ls})}{a_I} \int_{t_{bi}}^{t_I} dt e^{H_I(t_I - t)} = \frac{a(t_{ls})}{a_I H_I}(e^{N_e} - 1) \simeq \frac{a(t_{ls})}{a_I H_I} e^{N_e}. \tag{8.44}$$

This needs to be compared to the distance travelled by light from t_{ls} to now (t_0), measured at t_{ls}, i.e. ($H_0 = 2/(3t_0)$ during matter domination)

$$r_H(t_{ls}) = a(t_{ls}) \int_{t_{ls}}^{t_0} \frac{dt}{a(t)} = \frac{a(t_{ls})}{a_0} 3t_0 = \frac{2a(t_{ls})}{a_0 H_0}. \tag{8.45}$$

So the condition

$$\frac{d_H(t_{ls})}{2r_H(t_{ls})} > 1 \tag{8.46}$$

amounts to the same condition as obtained for the flatness problem,

$$e^{N_e} > \frac{a_I H_I}{a_0 H_0}. \tag{8.47}$$

Monopole Problem

As we said, the problem is that unlike other species, monopoles don't annihilate, and would remain until now. In a phase transition like the GUT phase transition, we expect to create about one monopole per nucleon. But from direct searches in the Universe, we know that there are less than 10^{-30} monopoles per nucleon. So there are too many monopoles (or in general, relics), and we need to dilute them somehow. It is very important that this dilution happens *after* the phase transition, so inflation needs to happen after the transition, or at most simultaneously with it.

The way the dilution happens is by the exponential expansion of a small causal patch. Since we have created one monopole per nucleon, we need expand the size of the Universe, defined by $a(t)$, by at least 10^{10}, so that the volume of the Universe grows by 10^{30}. Monopoles get diluted this way, and of course also do nucleons. But nucleons are created in much bigger numbers at the end of inflation, during reheating. So by diluting the monopoles (and other relics as well), we solve the problem. We need at last

$$N_e > ln 10^{10} = 23 \tag{8.48}$$

e-folds, which is a weaker bound than we found before. Of course, it actually says that the phase transition creating the relics needs to happen at least 23 e-folds before the end of inflation.

Why is the Entropy Large?

The reason the entropy per baryon is large is related to the presence of quantum fluctuations, and with reheating, that will populate the Universe with photons. This will be studied later on in the book. Since we have an exponential expansion, the size of the Universe will increase, leading to a large *total* number of photons.

Why Do We Have Perturbations?

We will explain this later, but it is basically because of quantum fluctuations, that will be exponentially blown up until they become classical and freeze out outside the horizon.

Why Do We Have a (Very Small) Baryon Asymmetry?

The baryon asymmetry itself is due to CP violation, but the conditions given by Sakharov (baryon number violation, CP violation, and interactions out of equilibrium) for the creation of baryons (baryogenesis) include the existence of interactions

out of equilibrium, satisfied by a fastly expanding Universe like the case of inflation. Moreover, the issue is also why there is such a *very small* baryon asymmetry. The answer is that we have a large entropy per baryon ($\sim 10^9$), that gets converted into baryon-antibaryon pairs, and leads to $(N_B - N_{\bar{B}})/N_B \sim 10^{-9}$.

8.4 Inflation with a Scalar Field

Finally, we describe the most popular incarnation of the inflationary paradigm, using an approximately constant energy density for a scalar field moving down a potential. The scalar field is called an inflaton, and its equations of motion in the FRLW Universe become the Friedmann equation

$$\left(\frac{\dot{a}}{a}\right)^2 + \frac{K}{a^2} = \frac{1}{3M_{\rm Pl}^2}\left[\frac{\dot{\phi}^2}{2} + V(\phi)\right], \tag{8.49}$$

and the Klein–Gordon equation,

$$\ddot{\phi} + 3\frac{\dot{a}}{a}\dot{\phi} = -V'(\phi). \tag{8.50}$$

Note that the Hubble expansion term (the one with $\dot{a}/a$) is a friction term, that moves against the acceleration of the scalar.

There is no a priori need to have a description in terms of a scalar field, it is just the simplest possibility, but logically, it can be also in terms of a vector, or tensor field, etc.

Old Inflation

The first model, by Alan Guth, called "old inflation", was based around the idea of a phase transition, with the idea in mind to solve the monopole problem, by having inflation in the same time, and continuing slightly after, as the phase transition.

A Higgs ("Mexican hat") potential, with a minimum at some $\phi_0 \neq 0$, and a nonzero value at $\phi = 0$ gets corrected by temperature effects in quantum field theory, so that for $T > T_c$, the true vacuum is at $\phi = 0$, for $V(0) > 0$, and ϕ_0 is lifted above it, as in Fig. 8.1. At $T = T_c$, the two vacua at 0 and ϕ_0 coincide. And at $T < T_c$, but not too small, there is a false vacuum at $\phi = 0$, separated from the true vacuum at ϕ_0 by a potential barrier.

Then the Universe starts in a hot state, at the (then) minimum of $\phi = 0$. As it cools off below T_c, it still remains there, now because of the potential barrier. Since we are at a constant $V(0) > 0$, the Universe undergoes a period of exponential inflation. But then it tunnels quantum mechanically (or jumps over due to thermal fluctuations) through the barrier, creating a small bubble.

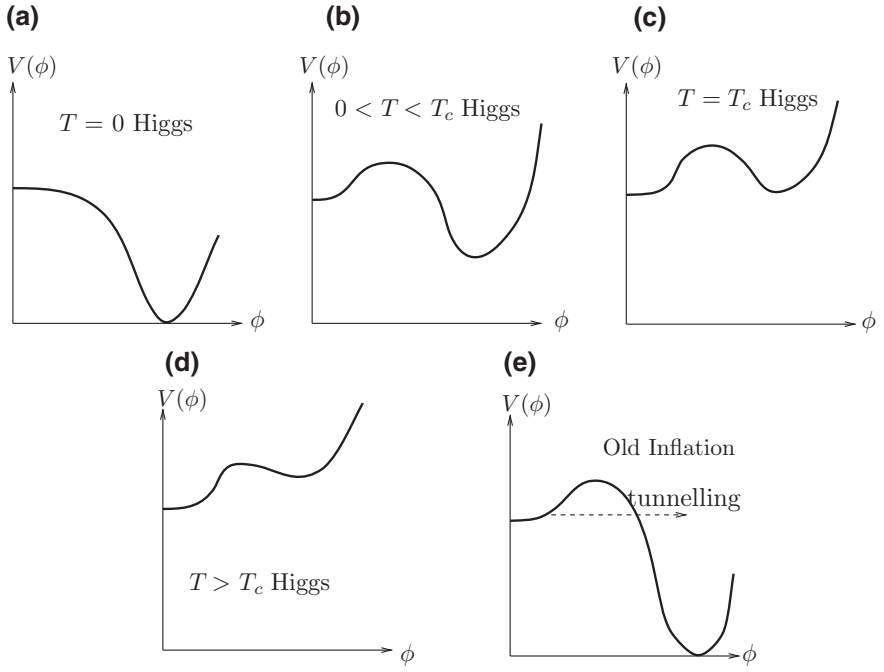

Fig. 8.1 **a** Higgs potential at $T = 0$. **b** Higgs potential at $0 < T < T_c$. **c** Higgs potential at $T = T_c$. **d** Higgs potential at $T > T_c$. **e** Old inflation potential is like case **b**, but has the true minimum at $V = 0$ (so is a combination of **a** and **b**)

The bubble, with a wall having a surface tension defined by $V(0)$, starts expanding, like a bubble of water in undercooled water vapor (or of vapor in overheated water), because it destabilizes the neighbouring regions into collapsing. This mechanism has been described by Coleman and DiLuccia. So it would seem that this bubble nucleation would give an end to inflation, since various bubbles would nucleate, and then eventually collide. This is what happens in water. But unlike water, in the inflationary Universe, the Universe in between the bubbles is still in the false vacuum at $\phi = 0$, so it still inflates, i.e., expands exponentially. That means that despite the fact that the walls of bubbles move at the speed of light, neighbouring bubbles never meet, since the space in between them expands exponentially. This is known as the "graceful exit problem" of old inflation.

Because of it, the "old inflation" model was dropped, and the "new inflation" of Albrecht, Steinhardt and Linde, supplanted it. In it, the scalar potential is almost flat, a plateau, but with a small slope, followed by a drop. Since Hubble expansion acts as a friction term, the movement on the plateau will be slow. This is the "slow-roll" paradigm, that will be described in the next chapter.

Important Concepts to Remember

- The smoothness and horizon problem is why is the Universe smooth and isotropic? Since the ratio of the distance travelled by light from last scattering until today, in today's scales, divided by the horizon at last scattering, in today's scales, is $N = 2r_H(t_0)/d_H(t_0) \sim 72$.
- The flatness problem is why has the Universe $\Omega \simeq 1$ to an absurd precision in the past (assuming $\Omega \sim 1$ today leads to $\Omega - 1 \sim 10^{-16}$ at electron-positron annihilation, for instance).
- Why is the total entropy in the Universe today of about 10^{88} (very large), why do we have a very small baryon asymmetry, how do we generate the perturbations that grow into structure (stars and galaxies); are also issues.
- The monopole (or relic) problem: what dilutes relics appearing in phase transitions, say (such that naively, we have as many baryons as relics).
- The inflationary paradigm requires most generally a period of accelerated expansion $\ddot{a} > 0$. The most common way is through a constant energy density, which leads to exponential expansion, $a \propto e^{Ht}$ (though power law inflation $a \propto t^p$, $p > 1$ is also fine), and usually it is assumed that it is parametrized by a scalar field, though it is likely not necessary, only simpler.
- Inflation happens near the Planck scale, but below it, so we can have an effective field theory description.
- The flatness problem is solved since $\Omega - 1 \propto t^{2(1-p)}$ or $\propto e^{-2Ht}$ drops very fast during inflation.
- Assuming that both now, and at the beginning of inflation, we have $\Omega \sim 1$, we find that we need at least $N_e > 56 + \ln \frac{[\rho_1]^{1/4}}{5\times 10^{13}\text{GeV}}$ e-folds of inflation.
- The smoothness and horizon problem is solved since during inflation, scales get blown up exponentially, while the horizon stays fixed, so scales get outside the horizon fast, and come back inside the horizon during standard cosmology. We obtain the same constraint as above on N_e.
- The monopole problem is solved if the phase transition occurs before or during inflation, so the generated relics of order one per baryon (in GUTs) are diluted away to less than 10^{-30} today, so need to be diluted by a factor of $e^{3N_e} > 10^{30}$, or $N_e > 23$ e-folds.
- Old inflation was inflation in a false vacuum $\phi = 0$, created as the temperature drops, whereas the true vacuum shifts to $\phi_0 \neq 0$. Then tunneling through the barrier ends inflation, but we have no "graceful exit", since bubbles of true vacuum never meet.
- New inflation is rolling down a plateau in order to inflate, after which there is a drop to a minimum.

Further reading: See [1, 10, 11].

Exercises

(1) If inflation occurs at temperature scale T_i and has 60 e-folds, calculate when the scale that got outside the horizon at the begining of inflation will cross inside it.
(2) Calculate the monopole/nucleon ratio that would require the same constraint on the number of e-folds as the one from reheating.
(3) Show that the equations of motion of the Einstein+scalar action reduce to

$$\left(\frac{\dot{a}}{a}\right)^2 + \frac{K}{a^2} = \frac{1}{3M_{\rm Pl}^2}\left(\frac{\dot{\phi}^2}{2} + V(\phi)\right)$$
$$\ddot{\phi} + 3\frac{\dot{a}}{a}\dot{\phi} = -V'(\phi). \tag{8.51}$$

(4) What are the equations of motion for the gravity field coupled to an antisymmetric tensor field $A_{\mu\nu}$ with uniform (space independent) field strength H_{ijk} (instead of a scalar) in the FLRW Universe?
(5) If inflation occurs at the GUT scale, and now $1 - \Omega \simeq 10^{-3}$, how many e-folds of inflation do we need in order to obtain the current Ω from a generic one at the beginning of inflation?

Chapter 9
Slow-Roll Inflation

In this chapter we will explore the simplest incarnation of inflation, namely the inflation with a scalar field that gives an approximate cosmological constant. We will see that we can quantify the condition into what is known as "slow-roll inflation". We will not address other types of inflation in this book, though we will explain what they can be.

As we said at the end of last chapter, after Guth's "old inflation" model, where the scalar sits at the false vacuum $\phi = 0$, with $V(0) > 0$ and then tunnels out of the barrier to the true vacuum at $\phi = \phi_0$, came the "new inflation" model. In "new inflation", there is a plateau in the potential, so $V(\phi)$ is large, but flat, after which there is a drop, as in Fig. 9.1. But this paradigm was much generalized afterwards, to the point that even the simplest thing we can think of, a polynomial potential $V = \lambda\phi^p$, fits into it, but we need the scalar to *slowly roll* down the potential. In the case of the polynomial potential, and other similar ones, we can have a condition on initial conditions (specifically, for the polynomial, that we start at large field ϕ). In fact, even slow-roll is not necessary for *inflation*, it is only necessary for *(almost) exponential inflation*. But in this chapter we will focus on this slow-roll inflation.

Variations of the new inflation with a plateau are "hilltop inflation", where we just start near the maximum of a potential (fine-tuned initial conditions); and "natural inflation", where we have a periodic, "axion-like" potential. Indeed, the axion of QCD is the one example we know of, of a periodic potential. In string theory, we have more "axions" with periodic potentials.

As we saw in the previous chapter, the equations of motion of the gravity+scalar system are: the Friedmann equation

$$H^2 + \frac{K}{a^2} \equiv \left(\frac{\dot{a}}{a}\right)^2 + \frac{K}{a^2} = \frac{1}{3M_{\rm Pl}^2}\left[\frac{\dot{\phi}^2}{2} + V\right]; \tag{9.1}$$

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_9

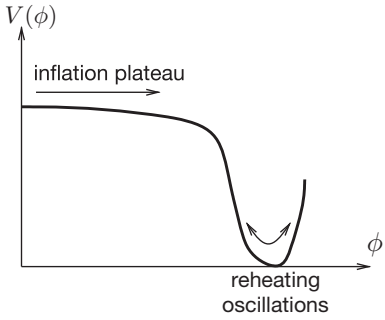

Fig. 9.1 The "new inflation" potential

note that here $\frac{\dot\phi^2}{2} + V(\phi) = \rho_\phi$ is the density of the scalar field, whereas the pressure is $P_\phi = \frac{\dot\phi^2}{2} - V(\phi)$. We will shortly neglect k/a^2, since this goes quickly to zero (in a couple of e-folds), due to the exponential inflation. The second equation is the KG equation, but it can also be obtained from the continuity equation, i.e., the conservation of energy, $\dot\rho = -3H(\rho + P)$, by substituting $\rho = \rho_\phi$ and $P = P_\phi$. In both cases, the result is

$$\ddot\phi + 3H\dot\phi + V'(\phi) = 0. \tag{9.2}$$

9.1 Slow-Roll Analysis

We start the analysis with the assumption that slow roll means that the Hubble parameter $H(t)$ has a relative variation in the time = horizon time $t_H = H^{-1}$ that is much less than one,

$$\frac{|\dot H|}{H^2} = -\frac{\dot H}{H^2} \equiv \epsilon_H \ll 1. \tag{9.3}$$

This defines the first slow-roll parameter ϵ_H. Taking the time derivative of (9.1), we get

$$2H\dot H = \frac{\dot\phi}{3M_{\rm Pl}^2}[\ddot\phi + V'] = -\frac{H\dot\phi^2}{M_{\rm Pl}^2}\ , \tag{9.4}$$

where in the second equality we have used (9.2). Then finally

$$\dot H = -\frac{\dot\phi^2}{2M_{\rm Pl}^2}. \tag{9.5}$$

Then the slow-roll condition $|\dot H| \ll H^2 = (\dot\phi^2/2 + V)/(3M_{\rm Pl}^2)$ becomes

$$\dot\phi^2 \ll |V(\phi)|\ , \tag{9.6}$$

i.e., that the kinetic energy of the scalar is much smaller than its potential energy. This is what we would have expected anyway for a slow-roll condition. In this case, we also obtain that $\rho \simeq -P \simeq V(\phi)$, which means that we have an *approximate cosmological constant*. Then as we saw in the previous chapter, the Hubble constant is

$$H \simeq \sqrt{\frac{V}{3M_{\rm Pl}^2}}. \tag{9.7}$$

The next condition for slow roll is similar to the first, but refers to the scalar field instead of H, namely that the relative variation of $\dot{\phi}$ during the horizon time $t_H = H^{-1}$ is small,

$$\left|\frac{\ddot{\phi}}{\dot{\phi}}\right| \frac{1}{H} \ll 1. \tag{9.8}$$

Since $|\ddot{\phi}| \ll |\dot{\phi}H|$, we can neglect the first term in the KG equation (9.2), so that it becomes

$$3H\dot{\phi} \simeq -V'(\phi)\ , \tag{9.9}$$

meaning that we can write

$$\dot{\phi} \simeq -\frac{V'}{3H} \simeq -\frac{V'}{\sqrt{3V}} M_{\rm Pl}\ , \tag{9.10}$$

where in the last equality we have used the Friedmann equation in the slow-roll form (9.7). Then from (9.5) and the above formula for $\dot{\phi}$, we have the condition

$$\epsilon_H = \frac{|\dot{H}|}{H^2} = \frac{3\dot{\phi}^2}{2V} \simeq \frac{M_{\rm Pl}^2}{2}\frac{V'^2}{V^2} \equiv \epsilon \ll 1\ , \tag{9.11}$$

where the equality $\epsilon_H \simeq \epsilon$ is only valid in the case both slow-roll conditions on ϕ are satisfied ($\dot{\phi}^2 \lll |V|$ and $|\ddot{\phi}| \ll |\dot{\phi}H|$). Here we have defined the parameter (depending only on the potential)

$$\epsilon \equiv \frac{M_{\rm Pl}^2}{2}\left(\frac{V'}{V}\right)^2. \tag{9.12}$$

Thus in the case of slow roll, $\epsilon \simeq \epsilon_H \ll 1$.

To find the second condition for slow-roll in terms of the potential only, we start from the KG equation when neglecting its the first term, $\dot{\phi} = -V'/(3H)$, and we take a time derivative, obtaining

$$\ddot{\phi} = -\frac{V''\dot{\phi}}{3H} + \frac{V'\dot{H}}{3H^2} = \frac{V''V'}{9H^2} - \frac{V'^3 M_{\rm Pl}^2}{6V^2}\ , \tag{9.13}$$

where in the second equality, we have used again $\dot{\phi} = -V'/(3H)$, as well as (9.5). If $\epsilon \ll 1$, from its definition it means that $V'^3 M_{\rm Pl}^2/V^2 \ll V'$, but since we want to neglect the first, acceleration, term ($\ddot{\phi}$) in the KG equation with respect to the second, Hubble friction, term, and then the equation is $3H\dot{\phi} = -V'$, it means that we want $\ddot{\phi} \ll V$, and then (9.13) implies we need also

$$|V''| \ll 9H^2 = \frac{3|V|}{M_{\rm Pl}^2}. \tag{9.14}$$

Finally, that gives the condition

$$|\eta| \ll 1 \ , \tag{9.15}$$

where we have defined the second slow-roll parameter,

$$\eta \equiv M_{\rm Pl}^2 \frac{V''}{V} \simeq \frac{V}{3H^2}. \tag{9.16}$$

In conclusion, the slow-roll conditions can be reduced to conditions on the potential itself, $\epsilon \ll 1$ and $\eta \ll 1$.

Variation of Slow Roll Quantities with the Number of e-Folds

We define the number of e-folds between two times t_1 and t_2 as before,

$$N = \ln \frac{a(t_2)}{a(t_1)} \ , \tag{9.17}$$

so that

$$dN = \frac{da}{a} = -Hdt. \tag{9.18}$$

Then the first relation for the variation with respect to the number of e-folds is for $\ln H$,

$$-\frac{d(\ln H)}{dN} = \frac{dH}{H^2 dt} = -\epsilon_H \simeq -\epsilon. \tag{9.19}$$

Similarly, we find the variation of $\ln \epsilon = 2\ln(V'/V) + \ln M_{\rm Pl}^2/2$,

$$-\frac{d(\ln \epsilon)}{dN} = \frac{d\epsilon}{\epsilon H dt} \simeq \frac{2\dot{\phi}}{H}\left(\frac{V''}{V'} - \frac{V'}{V}\right) = -\frac{2V''}{V}M_{\rm Pl}^2 + \frac{2V'^2}{V^2}M_{\rm Pl}^2 \ , \tag{9.20}$$

where in the last relation we have used the KG equation $3H\dot{\phi} \simeq -V'$ and the Friedmann equation $H^2 = V/(3M_{\rm Pl}^2)$, so that finally

$$-\frac{d(\ln \epsilon)}{dN} = 4\epsilon - 2\eta. \tag{9.21}$$

The last variation equation is for the variation of η with the number of e-folds. From the definition of η and N,

$$-\frac{d\eta}{dN} = M_{\rm Pl}^2 \frac{d}{Hdt}\left(\frac{V''}{V}\right) = \frac{M_{\rm Pl}^2\dot{\phi}}{H}\left(\frac{V'''}{V} - \frac{V''V'}{V^2}\right) = -\frac{M_{\rm Pl}^2}{V/M_{\rm Pl}^2}V'\left(\frac{V'''}{V} - \frac{V''V'}{V^2}\right)$$
$$= -\left(\frac{M_{\rm Pl}^4 V'''V'}{V^2}\right) + 2\left(\frac{M_{\rm Pl}^2 V''}{V}\right)\left(\frac{M_{\rm Pl}^2}{2}\frac{V'^2}{V^2}\right), \tag{9.22}$$

where in the third equality we used the Friedmann equation $\dot{\phi} = -V/(3H)$. Defining

$$\xi^2 \equiv \frac{M_{\rm Pl}^4 V'''V'}{V^2}, \tag{9.23}$$

we have finally

$$-\frac{d\eta}{dN} = -\xi^2 + 2\epsilon\eta. \tag{9.24}$$

9.2 General (Non-slow Roll) Formulas

We now derive formulas that are always valid, not just in the slow-roll case (so they would be valid even for a "fast roll", or rather non-slow roll, inflation).

First, we have seen already that we have the exact relation (9.5), derived from the Friedmann (9.1) and KG (9.2) equations, so $\dot{\phi}^2 = -2\dot{H}M_{\rm Pl}^2$. Consider next that we have the implicit relation $H = H(\phi)$, defined *on the trajectory solving the equations of motion*, via $H = H(t)$ and $\phi = \phi(t)$, so $\dot{H} = H'(\phi)\dot{\phi}$ ($H'(\phi) = \dot{H}/\dot{\phi}$). Then we find

$$\dot{\phi} = -2M_{\rm Pl}^2 H'(\phi). \tag{9.25}$$

Replacing this $\dot{\phi}$ in the Friedmann equation (9.1), we find

$$H^2(\phi) = \frac{1}{3M_{\rm Pl}^2}[2M_{\rm Pl}^4 H'^2(\phi) + V(\phi)] \Rightarrow$$
$$H'^2(\phi) - \frac{3}{2M_{\rm Pl}^2}H^2(\phi) = -\frac{V(\phi)}{2M_{\rm Pl}^4}. \tag{9.26}$$

This takes the forma of a *Hamilton–Jacobi equation for* $H(\phi)$ (understood like similar to the action function $S(\phi)$ in the classical mechanics formulation of the Hamilton–Jacobi equation).

Using the formula for $\dot{H}$ in (9.5) and the Friedmann equation (9.1), we find

$$\epsilon_H \equiv -\frac{\dot{H}}{H^2} = 3\frac{\dot{\phi}^2/2}{V + \dot{\phi}^2/2} = 3\frac{2M_{\rm Pl}^4 H'^2}{V + 2M_{\rm Pl}^4 H'^2} = 2M_{\rm Pl}^2\left(\frac{H'(\phi)}{H(\phi)}\right)^2, \tag{9.27}$$

where in the third equality we have used (9.25) in the numerator and denominator, and in the fourth equality we have used (9.26) in the denominator. The end result is similar to the definition of ϵ in terms of $V(\phi)$, just that now we have ϵ_H in terms of $H(\phi)$.

By analogy with the case of ϵ in terms of $V(\phi)$, we can define a quantity analogous to η from $V(\phi)$, namely

$$\eta_H \equiv 2M_{\rm Pl}^2 \frac{H''(\phi)}{H(\phi)}. \tag{9.28}$$

Then from (9.25), taking a time derivative, we find

$$\ddot{\phi} = -2M_{\rm Pl}^2 \dot{\phi} H''(\phi) \ , \tag{9.29}$$

so that

$$\eta_H = \frac{2M_{\rm Pl}^2 H''(\phi)}{H(\phi)} = -\frac{\ddot{\phi}}{\dot{\phi} H}. \tag{9.30}$$

On the other hand, from (9.25), we have

$$\epsilon_H = 2M_{\rm Pl}^2 \frac{H'^2}{H^2} = \frac{\dot{\phi}^2}{2M_{\rm Pl}^2 H^2}. \tag{9.31}$$

Taking a time derivative on this relation, we find

$$\dot{\epsilon}_H = \frac{\ddot{\phi}\dot{\phi}}{M_{\rm Pl} H^2} - \frac{\dot{\phi}^2 \dot{H}}{M_{\rm Pl}^2 H^3}. \tag{9.32}$$

In the second term, we use $\dot{H} = \dot{\phi} H' = -2M_{\rm Pl}^2 H'^2$ (from (9.25)), to obtain

$$\dot{\epsilon}_H = -\frac{\ddot{\phi}}{H\dot{\phi}} \left(-\frac{\dot{\phi}^2}{M_{\rm Pl}^2 H} \right) + \frac{\dot{\phi}^2}{H M_{\rm Pl}^2} \frac{2M_{\rm Pl}^2 H'^2}{H^2} = -2H\eta_H \epsilon_H + 2H\epsilon_H^2. \tag{9.33}$$

Solving for η_H, we find

$$\eta_H = \epsilon_H - \frac{\dot{\epsilon}_H}{2H\epsilon_H}. \tag{9.34}$$

From $\epsilon_H = -\dot{H}/H^2$, we obtain

$$\eta_H = -\frac{\dot{H}}{H^2} - \frac{1}{2H} \left(\frac{\ddot{H}}{\dot{H}} - 2\frac{\dot{H}}{H} \right) = -\frac{\ddot{H}}{2H\dot{H}}. \tag{9.35}$$

Finally, in the slow-roll case, as we saw, the Friedmann equation is $\dot{\phi} = -V'/(3H)$, so

$$\eta_H = -\frac{\ddot{\phi}}{\dot{\phi}H} = \frac{1}{H\dot{\phi}}\left(\frac{V''\dot{\phi}}{3H} + \frac{V'\dot{\phi}^2}{6M_{\rm Pl}^2H^2}\right) = \frac{V''}{3H^2} + \frac{V'\dot{\phi}}{6H^3M_{\rm Pl}^2} = M_{\rm Pl}^2\frac{V''}{V} - \frac{V'^2}{2\cdot 9H^4M_{\rm Pl}^2}\,, \tag{9.36}$$

where in the second equality we have used (9.5) and in the fourth equality we have used again $\dot{\phi} = -V'/(3H)$. Therefore, in the slow roll case, we obtain

$$\eta_H \simeq M_{\rm Pl}^2\frac{V''}{V} - \frac{M_{\rm Pl}^2}{2}\frac{V'^2}{V^2} = \eta - \epsilon. \tag{9.37}$$

Finally, we compute a formula for the number of e-folds between time t and the end of inflation $t_{\rm end}$,

$$N \equiv \ln\frac{a(t_{\rm end})}{a(t)} = \int_t^{t_{\rm end}} H dt = -\frac{1}{2M_{\rm Pl}^2}\int_\phi^{\phi_{\rm end}} d\phi\frac{H}{H'(\phi)} = -\frac{1}{M_{\rm Pl}}\int_\phi^{\phi_{\rm end}} d\phi\frac{1}{\sqrt{2\epsilon_H}}\,, \tag{9.38}$$

where we have used (9.25) in the form $dt = -d\phi/(2M_{\rm Pl}^2H'(\phi))$.

Note that this is an *exact* formula, independent of slow-roll. In the slow roll case, $\epsilon_H \simeq \epsilon$, so

$$N \simeq -\int_\phi^{\phi_{\rm end}} \frac{d\phi}{M_{\rm Pl}}\frac{1}{\sqrt{2\epsilon}}. \tag{9.39}$$

9.3 Slow Roll Conditions

The slow-roll conditions can be conditions on the parameters of the potential, like in the case of potentials with a plateau (new inflation). They can alternatively be conditions on the initial conditions for inflation. For instance, consider the power law potential

$$V = g\phi^\alpha\,, \tag{9.40}$$

where $g > 0$ and $\alpha > 1$ (the last condition is so that $V'' > 0$, but otherwise it can be dropped). Then

$$2\epsilon = \left(\frac{\alpha M_{\rm Pl}}{\phi}\right)^2; \quad \eta = \alpha(\alpha-1)\frac{M_{\rm Pl}^2}{\phi^2}\,, \tag{9.41}$$

so $\epsilon \ll 1$, $|\eta| \ll 1$ can be solved by either $\alpha \ll 1$ (a condition on the parameters, which is discarded if $\alpha > 1$), or by the initial condition $\phi_0/M_{\rm Pl} \gg 1$. But even that condition needs to be supplemented with a condition on parameters, since for the validity of the effective field theory of the inflation scalar, we need to have no quantum gravity corrections. That in turn translates into the condition that, at least, $V(\phi_0) \ll M_{\rm Pl}^4$, which means that

$$g \ll \frac{M_{\rm Pl}^4}{\phi_0^\alpha}. \tag{9.42}$$

But we really need to choose $\phi_0 \gg M_{\rm Pl}$ not only for the initial condition, but also for a large interval, since we need to have at least 60 or so e-folds of inflation. That turns out to impose that

$$g \ll 2 \times 10^{-5} \tag{9.43}$$

in the case $\alpha = 4$, that is the power most favoured by experimental data among the possible power laws.

9.4 Exact Solution for Inflation

In most cases of potentials used for inflation, we cannot find an exact solution for the time evolution. There is one notable exception of potential, that can be used for checking the various approximations used in this chapter. We will check several, leaving the others as exercises.

The potential giving an exact solution is the inverse exponential,

$$V = ge^{-\lambda\phi}. \tag{9.44}$$

Among the possible solutions to the equations of motion (Friedmann and KG), there is the attractor solution

$$\phi(t) = \frac{1}{\lambda} \ln\left[\frac{g\epsilon^2 t^2}{M_{\rm Pl}^2(3-\epsilon)}\right] \tag{9.45}$$

together with the scaling

$$a \propto t^{1/\epsilon} \tag{9.46}$$

and where

$$\epsilon \equiv \frac{\lambda^2 M_{\rm Pl}^2}{2}. \tag{9.47}$$

From these solutions, we find

$$\dot{\phi} = \frac{2}{\lambda t} \quad \text{and} \quad H = \frac{1}{\epsilon t} = \frac{2}{\lambda^2 M_{\rm Pl}^2 t}. \tag{9.48}$$

Then the Friedmann equation for $H^2 = 1/(\epsilon^2 t^2)$, is

$$H^2 = \frac{1}{3M_{\rm Pl}^2}\left(\frac{\dot{\phi}^2}{2} + V\right) = \frac{1}{3M_{\rm Pl}^2}\left(\frac{2}{\lambda^2 t^2} + g\frac{M_{\rm Pl}^2(3-\epsilon)}{g\epsilon^2 t^2}\right) = \frac{1}{\epsilon^2 t^2}\,, \tag{9.49}$$

so indeed it is satisfied. Also the KG equation,

$$\ddot{\phi} + 3H\dot{\phi} + V'(\phi) = -\frac{2}{\lambda t^2} + 3\frac{2}{\lambda t}\frac{1}{\epsilon t} - \frac{\lambda g M_{\rm Pl}^2(3-\epsilon)}{g\epsilon^2 t^2} = 0\,, \tag{9.50}$$

is satisfied.

The slow roll conditions give

$$\epsilon = \frac{M_{\rm Pl}^2}{2}\frac{V'^2}{V^2} = \frac{\lambda^2 M_{\rm Pl}^2}{2}; \quad \eta = M_{\rm Pl}^2\frac{V''}{V} = \lambda^2 M_{\rm Pl}^2 = 2\epsilon. \tag{9.51}$$

We note therefore that ϵ was well named, since indeed is the slow-roll parameter for the potential, and it is constant. The number of e-folds on the solution is

$$N = \ln\frac{a(t_2)}{a(t_1)} = \frac{1}{\epsilon}\ln\frac{t_2}{t_1} = \frac{\lambda}{2\epsilon}(\phi_2 - \phi_1) = \frac{1}{\lambda M_{\rm Pl}^2}(\phi_2 - \phi_1). \tag{9.52}$$

This matches with the result for a constant ϵ,

$$N \simeq -\frac{1}{\sqrt{2\epsilon}M_{\rm Pl}^2}(\phi_2 - \phi_1) = -\frac{1}{\lambda M_{\rm Pl}^2}(\phi_2 - \phi_1). \tag{9.53}$$

The slow-roll approximation for the solution is obtained neglecting $\ddot{\phi}$ in the KG equation, so that we need to solve $3H\dot{\phi} = -V'(\phi) = +g\lambda e^{-\lambda\phi}$, solved by

$$\phi(t) = \frac{1}{\lambda}\ln\left[\frac{\epsilon^2 t^2}{3M_{\rm Pl}^2}\right], \tag{9.54}$$

which is indeed the general solution, neglecting the ϵ in the $3-\epsilon$ factor in the denominator. The same solution also solves the Friedmann equation in which we drop the $\dot{\phi}^2/2$ term.

9.5 Variants of Inflation

Besides the general slow-roll paradigm for inflation, we have also other ideas that became important.

Eternal Inflation

This idea, originating with Steinhardt and others, is that, if we consider probabilities to be at different values of the potential V, distributed perhaps with a Boltzmann weight (or some other probability decreasing with energy), then the probability to be at large $V(\phi)$ will be small. If the probability refers to various patches of space to have some initial condition for the field, the initial patch with large V will be very small. However, since by being at $V \neq 0$ we inflate, the volume of the small patch quickly increases, also exponentially, so it could dominate.

Naively then, whether this patch will dominate or not depends on the relative strength of the coefficient β in the probability weight $e^{-\beta V}$ and the amount of exponential expansion of the volume $\mathcal{V} \sim e^{+3Ht} \sim e^{\sqrt{3V}t}$. But of course, in reality, we

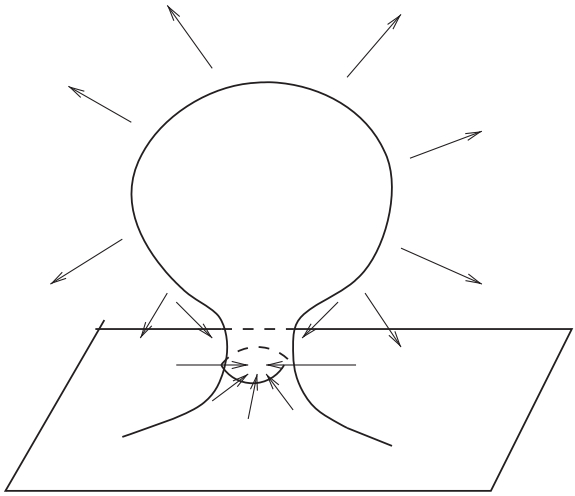

Fig. 9.2 A patch of higher V inside one of lower V creates a Baby Universe that peels off from the original one: there is a throat connecting the Baby Universe to the original one, that is collapsing. While the Baby Universe expands, viewed from the original one, the throat collapses

don't know what is the probability distribution over volumes, and moreover, it is not clear that we can even apply statistics to cosmology, where we have only one Universe (and not an ensemble of Universes; at least not to explore, there could be un-detectable ones in the many-worlds interpretation of Quantum Cosmology).

Moreover, we can consider fluctuations for the value of the potential: fluctuating upwards in V costs energy, so is less likely, but it has a higher benefit because of inflation. But again the issue of probabilities is a very tricky one.

So the issue of whether or not eternal inflation leads to an inflation defined by a large potential V seems a priori impossible to decide. What we do know is that in this idea, there will always be inflating regions, since there are always small patches of the Universe that have some high potential value, and these will eventually dominate over the rest. If they are outside the horizon, we cannot observe them, so the point is likely moot. And if they are inside the horizon, they might eventually overrun our own Universe, effectively killing it (though there is some debate about the efficiency of that scenario).

Chaotic Inflation

This idea, due to Linde, is that the initial conditions are random in space, so inflation occurs in patches. Moreover, when a patch inflates, it becomes homogenous and isotropic, smoothing out all initial inhomogeneities and anisotropies. The evolution of a patch of high $V = V_2$ inside one of lower V_1 is such that the patch inflates faster as viewed from inside it, but viewed from the outside, there is a "throat" that collapses due to its surface tension ($T = (V_2 - V_1)/Area$), as in Fig. 9.2. So eventually we have this Universe that detaches (its connection with the rest shrinks), but expands, smoothing out all inhomogeneities and vanishing an initial $\Omega - 1$.

The chaotic inflation picture is related to potentials without a plateau, but rather continuously related to zero, like the polynomial $V = g\phi^{\alpha}$, which is its standard

example. Having an initial condition at large ϕ is associated with a (quantum or thermal) fluctuation up the potential, like in the eternal inflation idea above.

Hybrid Models

We have also considered only single-scalar models until now, but that is not necessary. The simplest generalization is in the so-called hybrid models, which contain two scalars, one (σ) with its own potential $V(\sigma)$, and the other with a potential coupling depending on σ,

$$V(\phi, \sigma) = \frac{g^2}{2}\phi^2\sigma^2 + V(\sigma). \tag{9.55}$$

Multi-scalar Models

Finally, there could be multi-scalar models (N-flation, etc.), but we will not describe them. Motion in a multi-scalar potential landscape can help with some issues, but in most cases amounts to simply powers of N in various formulas.

9.5.1 Initial Conditions for Inflation

The issue of initial conditions is a tricky one, and depending on the point of view, can be considered a problem (e.g., see papers by Steinhardt in 2013), or not (e.g. papers by Linde and Guth in the same year). One point is that we needed to assume $\dot{\phi}^2/2 \ll V(\phi)$ at all times, including initially. The claim is that this is not generic, and equipartition of energy would demand that initially both terms are comparable. But this assumes a lot about the initial condition defined by quantum gravity (which could appear "out of nothing" like in the gravitational instanton defining the wave function of the Universe in a "no-boundary boundary condition" of Hartle and Hawking), as well as about quantum gravity itself (is equipartition of some *effective field theory* for the scalar reasonable?)

Other issues are more obvious: as we said, homogeneity and isotropy needs not to be assumed; in the chaotic inflation scenario, it is something that comes out of the dynamics, not initial conditions.

The last point to be addressed is of the initial condition for $V(\phi)$. We need that $V(\phi_0)^{1/4} \ll M_{\rm Pl}$ (slightly lower than the quantum gravity scale), since we need that the effective field theory of the scalar is a good approximation, meaning we are below the Planck scale, yet the initial condition is defined by quantum gravity, so we are close to it.

Important Concepts to Remember

- Slow-roll inflation requires $\epsilon_H = -\dot{H}/H^2 \ll 1$ ($\dot{\phi}^2 \ll V(\phi)$) and $\eta = M_{\rm Pl}^2 V''/V \ll 1$.
- During slow-roll, $\epsilon_H \simeq \epsilon \equiv M_{\rm Pl}^2/2(V'/V)^2$.
- In a general Hamilton–Jacobi picture, we have $\epsilon_H = 2M_{\rm Pl}^2(H'(\phi)/H(\phi))^2$, and we can define $\eta_H = 2M_{\rm Pl}^2(H''(\phi)/H(\phi))$.

- During slow-roll, $\eta_H \simeq \eta - \epsilon$.
- The number of e-folds during inflation is $N_e = -\int d\phi/M_{\rm Pl}/\sqrt{2\epsilon_H}$.
- The slow-roll conditions are conditions on the parameters and the initial conditions for the fields. Imposing also the number of e-folds, usually results in conditions on the parameters only.
- $V = ge^{-\lambda\phi}$ gives an exact solution for inflation, on which we can check the various formulas.
- Eternal inflation means that patches that have higher energy inflate faster, so can dominate the ensemble.
- Chaotic inflation means that we have random values for random patches, so somewhere always inflates, but on the average we roll down.
- There is a debate over the naturalness of the initial conditions for inflation.

Further reading: See [10, 11].

Exercises

(1) Consider the potential

$$V(\phi) = V_0\left(1 - ce^{-a\frac{\phi}{M_{\rm Pl}}}\right). \tag{9.56}$$

Calculate ϵ, η and the number of e-folds and put conditions on the parameters and initial conditions such that we have slow roll inflation.
(2) For $V(\phi) = ge^{-\lambda\phi}$ and the attractor solutions, calculate $\epsilon_H, \dot{\epsilon}_H$ and η_H, and verify

$$\eta_H = \epsilon_H - \frac{\dot{\epsilon}_H}{2H\epsilon_H}. \tag{9.57}$$

Compare also η_H with $\eta - \epsilon$ and check whether we have slow-roll inflation.
(3) For $V = g\phi^p$ calculate ϵ, η, N, $d\eta/dN$ and $d(\ln\epsilon)/dN$.
Calculate $a(t)$ in the slow-roll case and from it ϵ_H and η_H.
(4) Repeat Exercise 3 for $V = V_0[1 - c(\phi_0/\phi)^p]$, where $p > 0$.
(5) Consider scalar potential given by a cosine plus a constant,

$$V = A[1 + \cos(a\phi)]\,, \tag{9.58}$$

and inflation happening near the maximum at $\phi = 0$. Calculate ϵ, η and N_e, and find over what region we have slow roll, given the constants A, a and the initial value $\phi = \phi_0, \dot{\phi}_0 \ll 1/a$.

Chapter 10
Reheating and Baryogenesis

In this chapter we will study reheating, which happens just after inflation, and is a way to populate the Universe with thermally distributed particles, which is followed by the standard Radiation Dominated adiabatic cosmology. Thus reheating is a way to transfer the energy of the inflaton to the Standard Model particles and to transition into standard cosmology. We will also say a few things about baryogenesis which happens around the same time.

10.1 Standard Reheating

The simplest model of reheating happens in the standard models of "new inflation", with a plateau, followed by a drop to a minimum, and then oscillations. The inflaton field falls down to the minimum, then oscillates, until the friction terms stop it. During the oscillations, the inflaton transfers its energy by decaying into particle modes. In the model studied here, ϕ decays mostly into the fermionic modes, $\phi \to \psi\psi$, with a decay constant Γ_ϕ. This decay is a friction term in the equation of motion, or otherwise a relative friction term in the kinetic energy, $\dot{\rho}_\phi = -\Gamma_\phi \rho_\phi + ...$

The KG equation of motion of the scalar is modified to

$$\ddot{\phi} + 3H\dot{\phi} + \Gamma_\phi \dot{\phi} + V'(\phi) = 0. \tag{10.1}$$

In this equation of motion, besides the usual Hubble friction term $3H\dot{\phi}$, we have the explicit friction term due to the ϕ decay. The solution of the equations of motion corresponds to some coherent oscillations around the minimum (the field comes with some velocity, i.e., kinetic energy, and oscillates -damped- around the minimum). We multiply the modified KG equation with $\dot{\phi}$, and take the average over the oscillations (cycles), considering that the inflaton energy density,

$$\rho_\phi = \frac{\dot{\phi}^2}{2} + V \ , \tag{10.2}$$

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_10

gives on the average over a cycle (due to the equipartition of energy, kinetic = potential)

$$\langle \frac{\dot{\phi}^2}{2} \rangle + \langle V \rangle = \langle \dot{\phi}^2 \rangle. \tag{10.3}$$

We obtain then

$$\dot{\rho}_\phi + 3H\rho_\phi + \Gamma_\phi \rho_\phi = 0. \tag{10.4}$$

It is easy to see that the solution of this equation in terms of the energy density at the initial time t_I, corresponding to the end of inflation, when $\rho(t_I) = V(\phi_{\text{initial}}) \equiv M^4$ (defining the mass scale M), is

$$\rho_\phi(t) = M^4 \left(\frac{a_I}{a}\right)^3 e^{-\Gamma_\phi (t - t_I)}. \tag{10.5}$$

Next we write an equation for the matter (fermions) into which the inflaton decays. The conservation equation for matter in the absence of particle creation would have been the one considered in the first chapter, $\dot{\rho}_m + 3H(\rho_M + p_M) = 0$, where ρ_M is the density of matter (not inflaton), and p_M its pressure. But in fact, there is a driving term, which is minus the friction term of the inflaton, since the inflaton decays into matter. Therefore the equation of motion is

$$\dot{\rho}_m + 3H(\rho_M + p_M) = \Gamma_\phi \rho_\phi. \tag{10.6}$$

Since the mattter particles created during reheating (electrons, neutrinos, protons, etc.) are ultrarelativistic (as we saw in the thermal history of the Universe), $P_M = \rho_M/3$, and we have effectively radiation, therefore we rename $\rho_M \to \rho_R$, to obtain

$$\dot{\rho}_R + 4H\rho_R = \Gamma_\phi \rho_\phi. \tag{10.7}$$

We need to supplement this with the Friedmann equation

$$H^2 = \frac{1}{3M_{\text{Pl}}^2}(\rho_\phi + \rho_R). \tag{10.8}$$

Here we have used the fact that the total energy is composed of the energy of the inflaton and the radiation energy.

The solution of the equation of motion for ρ_R is

$$\rho_R(t) = \frac{\rho_\phi(t_I)\Gamma_\phi a^3(t_I)}{a^4(t)} \int_{t_I}^{t} dt'\, a(t') e^{-\Gamma_\phi (t' - t_I)}. \tag{10.9}$$

We can check that the $1/a^4(t)$ means that in $\dot{\rho}_R$ we get a $-4H\rho_R$ term, whereas the derivative acting on the integral gives

$$\frac{\rho_\phi(t_I)\Gamma_\phi a^3(t_I)}{a^4(t)} a(t) e^{-\Gamma_\phi(t'-t_I)} \ , \tag{10.10}$$

which is the same as Γ_ϕ times (10.5).

After it will thermalize, the energy density of radiation will be given by the standard relativistic thermal formula derived previously,

$$\rho_R = \frac{\pi^2}{30} g_* T^4 \ , \tag{10.11}$$

which allows us to calculate T out of ρ_R.

The initial time for the begining of reheating is the end of inflation, when the time is given by the Hubble time, $t = t_I \sim H^{-1} \sim M_{\rm Pl}/M^2$ (since by the Friedmann equation $H^2 \sim V/M_{\rm Pl}^2 \sim M^4/M_{\rm Pl}^2$). From this initial time until the time constant for the decay, $t \sim \Gamma_\phi^{-1}$, the inflaton will dominate the energy density (since it has not decayed yet). Because of its large mass m_ϕ, defined by the curvature of the potential around the minimum,

$$m_\phi^2 \equiv V''(\phi_0) \ , \tag{10.12}$$

ϕ is nonrelativistic, which means that the Universe is matter dominated, meaning that $a(t) \propto t^{2/3}$.

The decay occurs quantum mechanically, but from the tree level process, we can find the usual formulas

$$\Gamma_{\phi\to\psi\psi} = \frac{g^2 m_\phi}{8\pi} \ , \tag{10.13}$$

for the decay into fermions, and for the decay into bosons χ,

$$\Gamma_{\phi\to\chi\chi} = \frac{g^2}{8\pi m_\phi} \ , \tag{10.14}$$

where g is the coupling. We assume that $t_I \sim H^{-1} \ll \Gamma_\phi^{-1}$ (so that the above interval is nonzero), which means that $\Gamma_\phi \ll H(t_I)$, and therefore $e^{-\Gamma_\phi(t-t_I)} \simeq 1$.

We then obtain

$$\begin{aligned}
\rho_R(t) &\simeq \rho_\phi(t_I) \frac{\Gamma_\phi a^3(t_I)}{a^4(t)} \int_{t_I}^{t} dt'\, a(t') \\
&= \frac{3}{5}\Gamma_\phi \rho_\phi(t_I) t_I \left(\frac{t_I}{t}\right)^{8/3} \left[\left(\frac{t}{t_I}\right)^{5/3} - 1\right] \\
&= \frac{3}{5}\frac{\Gamma_\phi}{t}\rho_\phi(t_I) t_I^2 \left[1 - \left(\frac{t_I}{t}\right)^{5/3}\right] \\
&= \frac{3}{5}\frac{\Gamma_\phi}{t} M_{\rm Pl}^2 \left[1 - \left(\frac{t_I}{t}\right)^{5/3}\right] \ ,
\end{aligned} \tag{10.15}$$

where in the second equality we have used the fact that $a(t) \propto t^{2/3}$ during the period of matter domination (the massive, thus nonrelativistic, inflaton dominates the energy density), and in the last equality we have used the fact that $\rho_\phi(t_I) = M^4$ and $t_I^2 = M_{\rm Pl}^2/M^4$.

By taking the time derivative of this function, we obtain that it grows, from $\rho_R(t_I) = 0$ until a maximum, achieved at

$$t_{\rm max} = \left(\frac{8}{3}\right)^{3/5} t_I \ , \tag{10.16}$$

(which confirms that indeed $e^{-\Gamma(t_{\rm max}-t_I)} \simeq 1$ as was assumed) of value

$$\rho_{\rm R,max} = \left(\frac{3}{8}\right)^{8/5} \Gamma_\phi t_I \rho_\phi(t_I) \simeq 0.139 \Gamma_\phi \frac{M_{\rm Pl}}{M^2} M^4 = 0.139 \Gamma_\phi M_{\rm Pl} M^2. \tag{10.17}$$

After this maximum, $\rho_R(t)$ starts decreasing as

$$\begin{aligned} \rho_R(t) &\simeq \frac{3}{5}\Gamma_\phi M_{\rm Pl} M^2 \frac{t_I}{t}\left[1 - \left(\frac{t_I}{t}\right)^{5/3}\right] = \frac{3}{5}\Gamma_\phi M_{\rm Pl} M^2 \left(\frac{a_I}{a}\right)^{3/2}\left[1 - \left(\frac{a_I}{a}\right)^{5/2}\right] \\ &\sim \frac{3}{5}\Gamma_\phi M_{\rm Pl} M^2 (a_I/a)^{3/2} \ , \end{aligned} \tag{10.18}$$

where in the second equality we have used the fact that $a(t) \propto t^{3/2}$, and in the last one we have neglected the last term. We see then that, as long as we are still before $t \sim \Gamma_\phi^{-1}$, the radiation energy density drops as $a^{-3/2}$.

Finally, considering that this radiation energy density has already thermalized so that we can define a maximum temperature for the radiation, we use the relativistic relation to write

$$\rho_{\rm R,max} = \frac{\pi^2}{30} g_* T_{\rm max}^4 \ , \tag{10.19}$$

which gives the maximum temperature

$$T_{\rm max} \simeq 0.8 g_*^{-1/4} M^{1/2} (\Gamma_\phi M_{\rm Pl} \sqrt{8\pi})^{1/4}. \tag{10.20}$$

Entropy Production

In an *adiabatic* process, the total entropy $S = sa^3$ (where s is the entropy density per unit comoving volume) is conserved, so $s \propto 1/a^3$, and indeed, since in the relativistic case $s \propto T^3$ (the exact formula was $s = (2\pi^2/45) g_{*,s} T^3$), $s \propto 1/a^3$ ($T \propto 1/a$).

But in the case of reheating, we are out of equilibrium, so we are not in an adiabatic process. Entropy is in fact being generated, since the entropy density per comoving volume, $s \propto T^3$ as we saw, but also $\rho_R \propto T^4$, so $s \propto \rho_R^{3/4}$. Since $\rho_R \propto a^{-3/2}$ as we saw, we finally obtain

$$S \propto a^3 \rho_R^{3/4} \propto a^{15/8} \,, \tag{10.21}$$

so the entropy increases with the expansion of the Universe, as we said. In fact, we already said that producing the very large entropy of the Universe was one of the things that was needed of an inflationary model, and we do see that happening during reheating.

Finally, after $t \sim \Gamma_\phi^{-1}$, the inflaton ϕ has mostly decayed, and its energy has been converted into radiation, the total entropy levels off, and we enter the period of the standard adiabatic, radiation dominated cosmology.

Reheating Temperature

The most important quantity of reheating is the temperature at the end of reheating, when the standard cosmology starts, known as the *reheating temperature*. It is the temperature at $t = \Gamma_\phi^{-1}$, the last time that we can use the approximation of matter domination under which we have been doing the previous calculations, so

$$\begin{aligned} \rho_R(t = \Gamma_\phi^{-1}) &\simeq \frac{3}{5}\Gamma_\phi^2 \rho_\phi(t_I) t_I^2 = \frac{3}{5}\Gamma_\phi^2 M_{\rm Pl}^2 \\ &\equiv \frac{\pi^2}{30} g_* T_{RH}^4 \,, \end{aligned} \tag{10.22}$$

leading to

$$T_{RH} \equiv T(t = \Gamma_\phi^{-1}) \simeq 0.55 g_*^{-1/4} (\Gamma_\phi M_{\rm Pl} \sqrt{8\pi})^{1/2} \simeq g_*^{-1/4} \sqrt{\Gamma_\phi M_{\rm Pl}}. \tag{10.23}$$

On the other hand, from (10.13), for the fermionic decay $\phi \to \psi\psi$, we have $\Gamma_\phi \simeq (g^2/8\pi) m_\phi$, so

$$T_{RH} \simeq g_*^{-1/4} (g/\sqrt{8\pi}) \sqrt{m_\phi M_{\rm Pl}}. \tag{10.24}$$

10.2 New Reheating and Preheating

Nowadays however, another model of reheating is more popular. Instead of the many oscillations around a minimum, now it is assumed that the potential has a very steep drop, leading to what is known as *preheating*. This is an extremely rapid decay of modes into *bosons* (not fermions as before) via *parametric resonance*. In this case, there is no need for oscillations, or rather, oscillations stop as soon as they start (within less than a cycle). Because of the highly nonperturbative nature of this model, it is generally studied numerically, but here we will present a few salient points.

Generically, this new reheating picture is composed of 3 steps:

1. Preheating: the inflaton ϕ decays into bosons.
2. The bosons decay into everything.
3. The particles thermalize.

A simple model of the preheating step is obtained as follows. Consider the inflaton ϕ plus boson χ, with a potential composed of the mass term for ϕ (the curvature of the potential around the minimum) and a renormalizable (quartic) coupling of ϕ to χ, i.e.,

$$V \simeq \frac{1}{2}m_\phi^2\phi^2 + \frac{1}{2}g^2\phi^2\chi^2 + ... \tag{10.25}$$

Then we can effectively describe the field χ as having a *time-dependent mass* given by (the VEV of) ϕ,

$$m_\chi^2(t) \simeq g^2\phi^2(t). \tag{10.26}$$

The KG equation for the momentum modes χ_k of the boson χ becomes

$$\ddot{\chi}_k + E_k^2(t)\chi_k = 0\ , \tag{10.27}$$

where

$$E_k^2 = \left(\frac{k}{a}\right)^2 + m_\chi^2(t). \tag{10.28}$$

Assuming as before that the inflaton ϕ oscillates around its minimum, where the curvature is m_ϕ (even though as we will see we don't even need to complete one oscillation),

$$\phi = \phi_{\rm end}\sin(m_\phi t)\ , \tag{10.29}$$

where $\phi_{\rm end}$ is the value of ϕ at the end of inflation, we obtain for the χ mass

$$m_\chi^2(t) = g^2\phi_{\rm end}^2\sin^2(m_\phi t)\ , \tag{10.30}$$

and after defining rescaled variables and parameters as

$$z \equiv m_\phi t; \quad q \equiv \frac{g^2\phi_{\rm end}^2}{4m_\phi^2}; \quad A(k) \equiv \frac{k^2}{m_\phi^2} + 2q\ , \tag{10.31}$$

we find the equation

$$\frac{d^2\chi_k}{dz^2} + [A(k) - 2q\cos(2z)]\chi_k = 0. \tag{10.32}$$

This is the *Mathieu equation*. In certain regions of parameter space it has a rapidly growing solution, i.e., we have parametric resonance. In particular, the largest growth is for $q \gtrsim 1$. But a growth in the boson mode χ_k means particle production. Creating these bosonic particles amounts to preheating.

Numerically we can see that we can arrange for the parameters such that we don't even need to complete a full cycle in order to transfer the energy to χ. This is called *instant preheating*.

This process can be studied numerically, however one finds that the end result for the reheating temperature (the most important quantity characterizing reheating) is not changed much with respect to the standard reheating analysis before.

10.3 Entropy Production and the Number of e-Folds

We can now perform the more careful calculation of the bound on the number of e-folds coming from entropy production that was advertised in previous chapters.

Our Universe starts off as a patch at the begining of inflation of size = horizon size, $H^{-1} \sim M_{\rm Pl}/M^2$ (since $H^2 \sim V/M_{\rm Pl}^2$ by the Friedmann equation and $V \equiv M^4$). During inflation, this patch increases by a factor of e^N, where N is the number of e-folds. But the scale factor also increases during reheating. The increase in a is found from the fact that during reheating, as we saw, $\rho \propto 1/a^3$, and moreover at the end (when it has thermalized) we can write $\rho \sim T^4$, whereas at the begining of reheating we have $\rho \sim M^4$, so that finally

$$\frac{a_{RH}}{a_I} \simeq \left(\frac{M^4}{T_{RH}^4}\right)^{1/3} . \tag{10.33}$$

At the end of reheating, the entropy density $s = S/a^3$ is thermal and relativistic, so $s \sim T_{RH}^3$, so the total entropy within our patch of Universe is

$$S_{\rm patch} = a^3 s = \left(\frac{M^4}{T_{RH}^4} e^{3N} H^{-3}\right) T_{RH}^3 = e^{3N} \frac{M_{\rm Pl}^3}{T_{RH} M^2} . \tag{10.34}$$

But we already saw that the bound that solves the flatness problem is $S_{\rm patch} \geq 10^{88}$, so now we have

$$e^{3N} \frac{M_{\rm Pl}^3}{T_{RH} M^2} \geq 10^{88} , \tag{10.35}$$

which can be rewritten (by taking the log) as

$$N \geq N_{\rm min} = 67 + \frac{1}{3} \ln \frac{M^2 T_{RH}}{M_{\rm Pl}^3} , \tag{10.36}$$

where 67 comes from $\frac{88}{3} \ln 10$, or as

$$N \geq N_{\rm min} = 56 + \frac{2}{3} \ln \frac{M}{10^{16}\,{\rm GeV}} + \frac{1}{3} \ln \frac{T_{RH}}{10^9 {\rm GeV}} . \tag{10.37}$$

This was the more precise variant of the bound on the number of e-folds that was previously advertised.

10.4 Baryogenesis

Baryogenesis, the generation of the small baryon asymmetry observed in the Universe, must happen soon after, or at most during reheating. The reason is that, like in the case of primordial relics, any initial baryon asymmetry would be diluted away by inflation, so it must happen after it, i.e., during or after inflation. But it cannot happen too late, as we will shortly see.

In order to have baryogenesis, as we hinted at before, we must have satisfied the:

Sakharov Conditions

In fact, Andrei Sakharov (who, after being crucial in the Russian nuclear program during the Cold War, advocated for peace and got the Nobel Peace prize while being imprisoned by the state for his beliefs) proved that having baryogenesis is completely equivalent to having all the following 3 conditions satisfied:

1. *B breaking.* Baryon number (B) must be nonconserved in the process. This is in fact generic in Grand Unified Theories (GUTs), that unify electroweak ($SU(2) \times U(1)$) and strong ($SU(3)$) interactions in a common gauge group. Indeed, the very definition of unification of electroweak and strong forces imply the existence of "leptoquarks", i.e., gauge bosons from the unified gauge group that mediate transitions between quarks (transforming under $SU(3)$) and leptons (transforming under $SU(2)$), $q \leftrightarrow l$. But since baryons are by definition objects containing quarks, but no leptons, the leptoquarks violate B.
   B violation is clearly needed in order to obtain B asymmetry ($B \neq 0$) from $B = 0$.
2. *Breaking of C and CP symmetries.* In fact, C is violated in weak interactions, as we know (since in weak interactions CP is approximately conserved, and P is violated: this was the Nobel prize of T.D. Lee and C.N. Yang). Moreover, CP was found to be violated as well, in various sectors.
   This is less obvious to see, but it is also required.
3. *Absence of thermal equilibrium.* Indeed, in thermal equilibrium, the chemical potentials of particle and antiparticle are the same, $\mu_X = \mu_{\bar{X}}$, so we create an equal amount of particles and antiparticles. The rapid expansion of the Universe provides the mechanism of absence of thermal equilibrium. However, during standard adiabatic cosmology we do have thermal equilibrium, so the period of reheating (and shortly thereafter) is the one we need, since as we saw, then we do have non-equilibrium.

Bound on Temperature

Baryogenesis cannot happen too much after reheating also because it needs to be at a relatively high energy. Indeed, we saw that in order for it to happen, we need B violation, usually happening through GUT leptoquark decays. But there are many experiments looking for proton p decay; if found, it would mean that there are B violating leptoquark-like channels. But p decay has not been observed (which ruled out the simplest $SU(5)$ GUT for instance), and put very stringent bounds on the proton lifetime. Among other things, these bounds constrain the mass of the particles

mediating the decay to be $> 10^{10}$ GeV in generic models, though the bound can be smaller in specific (highly constrained) models.

In turn, this gives a constraint on the reheating temperature. Since baryogenesis can occur during reheating, this gives a constraint on T_{RH} that is slightly weaker,

$$T_{RH} \geq 10^9 \text{ GeV} \tag{10.38}$$

generically, though in very particular cases it can be even smaller.

10.5 Relics

Finally, a few words about relics, like monopoles, cosmic strings, etc. As we have seen, the necessity of diluting them was one of the motivations for inflation. These relics have usually very large masses (coming from phase transitions before, or during inflation), so during the standard RD era, beginning at the end of reheating, they are nonrelativistic, $\rho_{\text{relics}} \propto 1/a^3$, whereas the total energy density is dominated by radiation, so $\rho_{\text{tot}} \propto 1/a^4$, meaning that their ratio increases,

$$\frac{\rho_{\text{relics}}}{\rho_{\text{tot}}} = \frac{\rho_{\text{non-rel}}}{\rho_{\text{rel}}} \propto a \propto \frac{1}{T}. \tag{10.39}$$

Since radiation domination ends at matter-radiation equality, when the temperature is T_{eq}, and after that we have matter domination, so the relics and the total energy density scale in the same way, thus their ratio stays constant until now, from the reheating time until now the above ratio of energy densities changes by a factor of ($T_{eq} \sim 10^5 K \sim 10$ eV)

$$\frac{T_{RH}}{T_{eq}} = 10^{18} \times \left(\frac{T_{RH}}{10^{10} \text{ GeV}} \right). \tag{10.40}$$

That means that we need to suppress the relics by at least this factor during inflation (so relic production must happen at a number of e-folds generating at least the above factor from the end of inflation).

Important Concepts to Remember

- In standard reheating, at the end of inflation, the potential drops, and there is a minimum. There are many oscillations around the minimum. The inflaton energy is transmitted to fermionic modes, through a coupling $\phi\psi\psi$ giving decay $\Gamma_{\phi\to\psi\psi}$.
- Matter is ultrarelativistic fermions here, meaning radiation. Its density increasing to a maximum, corresponding to a maximum temperature after thermalization.
- Entropy is produced during reheating, $S \propto a^{15/8}$, but at its end we move on to adiabatic RD cosmology, with constant S.

- The reheating temperature T_{RH} is the most important quantity, which is the temperature of the radiation at $t = \Gamma_\phi^{-1}$, when the inflaton has stopped decaying.
- We obtain $T_{RH} \sim g_*^{-1/4}\sqrt{\Gamma_\phi M_{\rm Pl}}$.
- In new reheating, there is a period of preheating, which is a steep drop in the potential, during which the inflaton decays rapidly into bosons. Then the bosons decay into everything and we have thermalization. Preheating requires less than one oscillation cycle to occur.
- Preheating leads to the Mathieu equation, which has a parametric resonance in some region in parameter space.
- Correcting the number of e-folds of inflation for preheating, we find $N \geq 56 + \frac{2}{3}\ln\frac{\rho_1^{1/4}}{10^{16}\text{GeV}} + \frac{1}{3}\ln\frac{T_{RH}}{10^9\text{GeV}}$.
- The Sakharov conditions for baryogenesis are: 1. B breaking. 2. C and CP breaking. 3. Out of thermal equilibrium.
- Baryogenesis needs to occur during or not long after reheating, which gives generically (due to constraint on generic baryogenesis models) $T_{RH} > 10^9\,\text{GeV}$.
- Relics need to be diluted by a factor of $10^{18}T_{RH}/10^{10}\,\text{GeV}$ during inflation.

Further reading: See Chap. 8.3 in [1], Chap. 5.5 in [2] and Chap. 21 in [9].

Exercises

(1) Estimate the reheat temperature in the MSSM plus scalar = inflaton of mass $m \sim 10^{13}\,\text{GeV}$ coupled to the fermions.
(2) Estimate the minimum number of e-folds needed for inflation at the scale $10^{16}\,\text{GeV}$ in the case at Exercise 1.
(3) Solve the Mathieu equation as a perturbation (to first order) around a given time.
(4) Consider an inflationary plateau at GUT scale, followed by a rapid drop to a minimum. Assuming that the potential energy is quickly converted completely to thermalized radiation, and that at that energy scale we have only MSSM particles being effectively massless, calculate the reheating temperature.

Chapter 11
Fluctuations in Inflation and Matching with Experimental Data

In this chapter we will show how to calculate the spectrum of fluctuations during inflation, and relate that to observables. We will quantize fluctuations during inflation, and show what power spectrum they give, and how that reflects into the CMBR fluctuations. The final point will be how to then constrain various inflationary models using experimental data.

11.1 Scales During Inflation

As we saw, during inflation, scales get out of the horizon (which is approximately constant, $H^{-1} \simeq$ const.), being exponentially blown up, and then during the standard (RD, and later, MD) cosmology, they come back inside the horizon. Moreover, as we will see, perturbations start off as quantum fluctuations, but then when they get outside the horizon, they get "frozen in" and grow with a until they become classical, and when they come back inside the horizon have therefore classical, observable, effects. When speaking about scales, we will refer to wave number modes k in coordinate (x, not $a(t)x$) space, so that the physical (momentum) scales are actually $k/(a(t))$. Therefore the horizon exit will happen when $k/a = H$, or $k = aH$. Note that the scales that leave the horizon earliest spend most time outside the horizon, thus came back latest ("first out, last in").

The largest scales that are relevant for us today are the scales of the CMBR, or the Hubble radius, of the order of $\lambda \sim 3000 Mpc$. Any bigger scales are still outside the horizon, thus unobservable, thus irrelevant. Thus this largest scale left the horizon at the earliest time, corresponding to the minimum number of e-folds before the end of inflation. We have calculated this number of e-folds in the last chapter, so we can now say that

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_11

$$N_{3000Mpc} = N_{\min} = 56 + \frac{2}{3}\ln\frac{\rho_*^{1/4}}{10^{16}\,\mathrm{GeV}} + \frac{1}{3}\ln\frac{T_{RH}}{10^{9}\,\mathrm{GeV}}. \tag{11.1}$$

On the other hand, now we can calculate the number of e-folds before the end of inflation that an arbitrary scale λ has left, considering that $\ln 3000 \simeq 8$, namely

$$N_\lambda = 48 + \ln\frac{\lambda}{Mpc} + \frac{2}{3}\ln\frac{\rho_*^{1/4}}{10^{16}\,\mathrm{GeV}} + \frac{1}{3}\ln\frac{T_{RH}}{10^{9}\,\mathrm{GeV}}. \tag{11.2}$$

11.2 Scalar Fluctuations During Inflation

The simplest model of inflation, which we analyze here, is characterized by the interaction of gravity with a scalar field. Of course, we can have other fields, but this simplest model assumes that always the leading contribution of all the fields (other than gravity) will be described by a scalar.

To calculate the scalar fluctuations during inflation, we write the KG equation for a scalar with potential $V(\phi)$ and vary it with respect to a $\delta\phi$, after which we go to momentum space, obtaining the equation for $\delta\phi_k$,

$$\delta\ddot{\phi}_k + 3H\delta\dot{\phi}_k + \left(\frac{k}{a}\right)^2\delta\phi_k + V''(\phi)\delta\phi_k = 0. \tag{11.3}$$

Since during inflation we assume that the potential is very flat (so that we have slow-roll), we consider that $V''(\phi) \simeq 0$, and neglect it. Of course, in the previous chapter, when analyzing reheating, we have considered a very large inflaton mass, but that was given by the curvature of the potential around the minimum where reheating happens. On the plateau where inflation happens, the inflaton is nearly massless. The equation above is in the absence of backreaction, i.e., we consider that the scalar perturbation doesn't influence the gravitational background.

But in reality, there are gravitational perturbations as well. We have seen that we can choose a Newtonian gauge, where the metric perturbations are $h_{00} = -2\Psi$, $h_{0i} = 0$ and $h_{ij} = -a^2\delta_{ij}\Phi$, and $\Psi = \Phi$ for the scalar perturbations which we study here. Inputting these perturbations into the Einstein equation, $G_{\mu\nu} = 8\pi G_N T_{\mu\nu}$, and perturbing to first order, we obtain

$$\ddot{\Psi}_k + H\dot{\Psi}_k = 4\pi G_N\dot{\phi}\delta\phi_k\ , \tag{11.4}$$

since the Einstein tensor $G_{\mu\nu}$ is quadratic in derivatives. But that means that the scalar fluctuation $\delta\phi_k$ generates a Newtonian perturbation Ψ_k. That means that in reality we don't have a purely scalar perturbation, but rather a combination of the scalar and gravity modes. However, for now we will stick with the assumption of no backreaction, and later we will see that we need only modify it slightly for the real case.

We first redefine the scalar variable as

$$\varphi \equiv a\delta\phi \,, \tag{11.5}$$

and use conformal time η, defined by $d\eta = dt/a(t)$, instead of normal time t. Since $H \simeq$ constant during inflation, we have

$$\eta \simeq -\frac{1}{aH}. \tag{11.6}$$

Indeed, when differentiating, we find $d\eta = dt\dot{a}/(a^2H) = dt/a$. Then, rewriting the KG equation (11.3) at $V''(\phi) = 0$ in these variables, we obtain

$$\frac{d^2\varphi_k(\eta)}{d\eta^2} + \omega_k^2(\eta)\varphi_k(\eta) = 0 \,, \tag{11.7}$$

that is, a free harmonic oscillator, but with a time-dependent frequency,

$$\omega_k^2(\eta) = k^2 - \frac{2}{\eta^2} = k^2 - (aH)^2. \tag{11.8}$$

We indeed see that the moment of horizon exit, $k = aH$, delimitates real from imaginary frequency of oscillation:

1. For solutions on super-horizon scales, $k \leq aH$, or rather $k \ll aH$, implying $(-k\eta) \ll 1$, i.e., at late times, the last term in (11.3) (when $V''(\phi) \simeq 0$ also) is negligible, so the equation becomes

$$\delta\ddot{\phi}_k + 3H\dot{\phi}_k = 0 \,, \tag{11.9}$$

which has the solution $\delta\phi_k =$ constant. That means that once the solution gets outside the horizon, it gets "frozen in" at a constant value, and its size (scale) expands with $a(t)$. Eventually it will become classical, and will come back inside the horizon during the standard (RD or MD) cosmology. Once it gets back inside the horizon, it will start to fluctuate again, but this time as a classical solution.

2. For solution on sub-horizon scales, $k > aH$, or more precisely $k \gg aH$, implying $(-k\eta) \gg 1$, or early times, we have a simple harmonic oscillator with constant $\omega_k^2(\eta) \simeq k^2$, so we have the usual oscillating solution,

$$\varphi_k(\eta) \simeq \frac{e^{-ik\eta}}{\sqrt{2k}} \,, \tag{11.10}$$

where we have chosen a particular normalization.

The general solution in momentum space is

$$\hat{\varphi}_k(\eta) = \frac{1}{(2\pi)^{3/2}}[\varphi_k(\eta)\hat{a}(\vec{k}) + \varphi_k^*(\eta)\hat{a}^\dagger(-\vec{k})]. \tag{11.11}$$

At the classical level, $\hat{a}(\vec{k})$, $\hat{a}^\dagger(\vec{k})$ are coefficient functions for the modes, whereas as usual, at the quantum level they will turn into annihilation and creation operators for the modes. The solution in position space is simply the momentum integral of the above,

$$\delta\varphi(\vec{x}, t) = \frac{1}{(2\pi)^{3/2}} \int d^3k[\varphi_k(\eta)e^{i\vec{k}\cdot\vec{x}}\hat{a}(\vec{k}) + \varphi_{-k}^*(\eta)e^{-i\vec{k}\cdot\vec{x}}\hat{a}^\dagger(\vec{k})] \tag{11.12}$$

We consider the solution with the initial condition (at $\eta \to -\infty$) given by the oscillating solution,

$$\varphi_k(\eta) \to \frac{e^{-ik\eta}}{\sqrt{2k}}. \tag{11.13}$$

Then we can check explicitly (by calculating $d^2\varphi_k/d\eta^2$ and $\omega_k^2\varphi_k$) that

$$\varphi_k(\eta) = \frac{e^{-ik\eta}}{\sqrt{2k}} \frac{k\eta - i}{k\eta} \tag{11.14}$$

is a solution satisfying the initial condition.

Well after horizon exit, at $(-k\eta) \ll 1$, we have

$$\varphi_k(\eta) \to \frac{-i}{\sqrt{2k}k\eta}\ , \tag{11.15}$$

and the physical fluctuation is

$$\delta\phi_k = \frac{\varphi_k}{a} = \frac{+i/\sqrt{2k}}{k(-a\eta)} = \frac{iH}{\sqrt{2k}k} = \text{const.}(\eta). \tag{11.16}$$

Thus indeed the solution becomes constant when reaching super-horizon scales.

11.3 Quantization of Scalar Perturbations

The solution above starts off as a quantum fluctuation, and only after horizon exit it becomes classical. That means that we need to quantize the system, done as usual by promoting $\hat{a}(\vec{k})$ and $\hat{a}^\dagger(\vec{k})$ to operators, obeying the usual $[a, a^\dagger] = 1$ algebra, or rather, with continuous $\vec{k}$,

$$[\hat{a}(\vec{k}), \hat{a}^\dagger(\vec{k}')] = (2\pi)^3\delta(\vec{k} + \vec{k}'). \tag{11.17}$$

But if we are in quantum mechanics, we need to also specify the *state* that the Universe (or rather, the patch that becomes our visible Universe) is in. There are several possible choices, but the most common one is the simplest one, called the *Bunch-Davies vacuum* $|0\rangle$, defined as the state annihilated by $\hat{a}(\vec{k})$,

$$\hat{a}(\vec{k})|0\rangle = 0, \quad \forall \vec{k}. \tag{11.18}$$

The choice of this state corresponds in the classical picture to the intial condition (11.13) for the classical solution.

Power Spectrum

Choosing the Bunch-Davies vacuum state, and with the quantum field given by (11.11), we easily obtain that

$$\begin{aligned} \langle \varphi_k \rangle &\equiv \langle 0|\hat{\varphi}_k(\eta)|\rangle = 0 \\ \langle \varphi_k \varphi_{k'} \rangle &= \langle 0|\hat{\varphi}_k(\eta)\hat{\varphi}_{k'}(\eta)|0\rangle = |\varphi_k(\eta)|^2 \delta^3(\vec{k}+\vec{k}') \ , \end{aligned} \tag{11.19}$$

by using $\hat{a}(\vec{k})|0\rangle = 0 = \langle 0|\hat{a}^\dagger(\vec{k})$ and $\langle 0|\hat{a}(\vec{k})\hat{a}^\dagger(\vec{k}')|0\rangle = (2\pi)^3\delta^3(\vec{k}+\vec{k}')$.

But in general the power spectrum $\mathcal{P}_\varphi(k)$ is defined by

$$\langle \varphi_{\vec{k}} \varphi_{\vec{k}'} \rangle \equiv \frac{2\pi^2}{k^3} \mathcal{P}_\varphi(k) \delta^3(\vec{k}+\vec{k}') \ , \tag{11.20}$$

so we obtain

$$\mathcal{P}_\varphi(k,\eta) = \frac{k^3}{2\pi^2} |\varphi_k(\eta)|^2. \tag{11.21}$$

We are interested in the result after horizon exit, which then gets frozen in until today, when we can observe it in the CMBR, on the largest scales, that have just come back inside the horizon. Then we obtain

$$\mathcal{P}_\varphi(k,\eta) \to \frac{1}{(2\pi)^2 \eta^2}. \tag{11.22}$$

Since $\varphi_k = a\delta\phi_k$, we have

$$\langle \varphi_k \varphi_{k'} \rangle = a^2 \langle \delta\phi_k \delta\phi_{k'} \rangle \ , \tag{11.23}$$

so the power spectrum for the observable fluctuations $\delta\phi_k$ is

$$\mathcal{P}_\phi = \frac{1}{a^2} \mathcal{P}_\varphi = \left(\frac{H}{2\pi} \right)^2 \ , \tag{11.24}$$

where we have used the fact that $a\eta = -1/H$.

In momentum space, considering that $H \simeq H_* = \text{constant}$, and that the spectrum is approximately zero for sub-horizon scales, $k \geq aH_*$, and that the Universe is considered to have a size L, so that the minimum k is L^{-1}, is

$$\langle \delta\phi^2(\vec{x}, t) \rangle \simeq \left(\frac{H_*}{2\pi}\right)^2 \int_{L^{-1}}^{aH_*} \frac{dk}{k} = \left(\frac{H_*}{2\pi}\right)^2 \ln(LH_*a) = \left(\frac{H_*}{2\pi}\right)^2 N(t) \ , \quad (11.25)$$

where $N(t)$ is the number of e-folds after the Universe size aL leaves the horizon.

11.4 Scalar-Gravity Fluctuations

Until now we have considered only the fluctuations of the scalar, ignoring the back-reaction of gravity onto it. But as we saw, $\delta\phi$ generates a Newtonian potential fluctuation $\delta\Psi$, which in turn generates an extra contribution to $\delta\phi$ (through the perturbed metric in the $\Box$ operator), appearing on the right-hand side of the KG equation. We will not do this calculation here, but the result is (again for the case $V''(\phi) \simeq 0$)

$$\delta\ddot{\phi}_k + 3H\delta\dot{\phi}_k + \left(\frac{k}{a}\right)^2 \delta\phi_k = \left[\frac{1}{a^3}\frac{d}{dt}\left(\frac{a^3\dot{\phi}^2}{H}\right)\right]\delta\phi_k. \quad (11.26)$$

One can now also write the equation in terms of φ_k and η like in the pure scalar case, and obtain

$$\frac{d^2\varphi_k(\eta)}{d\eta^2} + \left(k^2 - \frac{1}{z}\frac{d^2z}{d\eta^2}\right)\varphi_k(\eta) = 0 \ , \quad (11.27)$$

where we have defined the variable

$$z \equiv \frac{a\dot{\phi}}{H}. \quad (11.28)$$

This is called the *Mukhanov-Sasaki equation*.

We have shown how to obtain the Mukhanov-Sasaki equation from the backreaction of gravity, in a specific gauge (the Newtonian gauge) onto the scalar. But of course, we can write the same equation in terms of a gauge-invariant variable, the curvature perturbation ζ_k introduced in Chap. 7 when discussing the CMBR. Of course, the gauge invariant formalism is more complicated, so obtaining the equation for ζ_k in it would be hard, but we will see how to use a shortcut.

The curvature perturbation can be understood at a non-rigorous level from a modification of the Friedmann equation due to the perturbation ζ,

$$H^2 = \frac{8\pi G_N}{3}\rho + \frac{2}{3}\nabla^2\zeta(\vec{x}, t) \ , \quad (11.29)$$

which when varied over a momentum scale k gives

$$2H\delta H_k = \frac{8\pi G_N}{3}\delta\rho_k - \frac{2}{3}\left(\frac{k}{a}\right)^2 \zeta_k. \tag{11.30}$$

But a more rigorous definition can be used. We have seen in Chap. 7 that for a general slicing, with

$$g_{ij} = a^2(t)e^{2\psi(\vec{x},t)}\gamma_{ij} \ , \tag{11.31}$$

with $\det\gamma_{ij} = 1$, we have $\psi = \zeta - H\delta t(\vec{x}, t)$, where the transformation $\delta t(\vec{x}, t)$ relates this gauge to the *uniform density slicing* where one defines ζ. Consider now that the general slicing is a *flat slicing*, with $\psi = 0$ (such that there is no curvature perturbation). On such a slicing, we can define the scalar fluctuation $\delta\phi(\vec{x}, t)$, related to δt by the obvious relation

$$\delta\phi(\vec{x}, t) = -\dot{\phi}\delta t(\vec{x}, t) \ , \tag{11.32}$$

since in the uniform density slicing there is no $\delta\phi$ and no δt. Then moreover, $\psi = 0$ implies

$$\zeta = H\delta t = -H\frac{\delta\phi}{\dot{\phi}}. \tag{11.33}$$

Using the definition of the quantity z, we then can write

$$\zeta = -\frac{a\delta\phi}{z} = -\frac{\varphi}{z}. \tag{11.34}$$

Therefore $\varphi = -z\zeta$, which means that the correct equation for the gauge invariant quantity ζ can be obtained by replacing in the Mukhanov-Sasaki equation (11.27) $\varphi = -z\zeta$.

Doing the replacement, we see that the term with $d^2z/d\eta^2$ cancels, and we are left with the equation for ζ_k,

$$\frac{d^2\zeta_k}{d\eta^2} + \frac{2}{z}\frac{dz}{d\eta}\frac{d\zeta_k}{d\eta} + k^2\zeta_k = 0. \tag{11.35}$$

In order to write it explicitly, we need to compute $dz/zd\eta$. To do so, we first consider the definition of z, $z = a\dot{\phi}/H$, from which we obtain

$$\frac{1}{z}\frac{dz}{d\eta} = \frac{a}{z}\frac{dz}{dt} = aH + a\frac{\ddot{\phi}}{\dot{\phi}} - \frac{a\dot{H}}{H}. \tag{11.36}$$

We have seen in Chap. 9 that $\dot{H} = -\dot{\phi}^2/(2M_{\rm Pl}^2)$, so we obtain

$$\frac{1}{z}\frac{dz}{d\eta} = aH\left(1 + \frac{\ddot{H}}{2H\dot{H}} - \frac{\dot{H}}{H}\right) = aH\left(1 + \epsilon_H + \eta_H\right) \ , \tag{11.37}$$

where we have used the definitions of the general slow-roll parameters $\eta_H = \ddot{H}/(2H\dot{H})$ and $\epsilon_H = -\dot{H}/H^2$. In order to express aH as well, we notice that

$$\frac{d}{d\eta}\left(\frac{1}{aH}\right) = a\frac{d}{dt}\left(\frac{1}{aH}\right) = -1 - \frac{\dot{H}}{H^2} = -1 + \epsilon_H. \tag{11.38}$$

The solution of this equation is then

$$aH = -\frac{1}{(1-\epsilon_H)\eta}. \tag{11.39}$$

During slow roll, $\epsilon_H \simeq \epsilon \ll 1$ and $\eta_H \simeq \tilde{\eta} - \epsilon$, where we have used $\tilde{\eta}$ for the slow-roll parameter denoted η previously, in order not to confuse with the conformal time. Finally, we find

$$aH \simeq -\frac{1+\epsilon}{\eta}\,, \tag{11.40}$$

and

$$\frac{1}{z}\frac{dz}{d\eta} \simeq aH(1+\tilde{\eta}) \simeq -\frac{(1+\epsilon)(1+\tilde{\eta})}{\eta} \simeq -\frac{1+\tilde{\eta}+\epsilon}{\eta}. \tag{11.41}$$

Substituting in the Mukhanov-Sasaki equation for ζ_k, (11.35), we find, in the case of slow-roll,

$$\frac{d^2\zeta_k}{d\eta^2} - \frac{2}{\eta}(1+\epsilon+\tilde{\eta})\frac{d\zeta_k}{d\eta} + k^2\zeta_k = 0. \tag{11.42}$$

The solution of this equation is

$$\zeta_k = -\frac{\sqrt{\pi}}{2(e\pi)^{3/2}}\frac{\sqrt{\eta}}{z(\eta)}e^{\frac{i\pi\nu}{2}+\frac{i\pi}{4}}H_\nu^{(1)}(-k\eta)\,, \tag{11.43}$$

where $H_\nu^{(1)}$ is the Hankel function and

$$\nu \equiv \frac{3}{2} + 2\epsilon_H + \eta_H = \frac{3}{2} + \epsilon + \tilde{\eta}. \tag{11.44}$$

The proof of this is left as an exercise. Note that the factor $1/z(\eta)$ appears since $z(\eta)\zeta_k = \varphi_k$ is the solution of the Mukhanov-Sasaki equation (11.27).

At late times when reaching super-horizon scales, $k \ll aH$ or $(-k\eta) \ll 1$, the solution becomes

$$\zeta_k \propto k^{-\nu} \simeq \text{const.}(\eta). \tag{11.45}$$

11.5 Power Spectrum and Contact with Experimental Data

Since $\zeta \simeq -H\delta\phi/\dot{\phi}$ as we saw, it means that

$$\langle \zeta_k \zeta_{k'} \rangle \simeq \frac{H^2}{\dot{\phi}^2} \langle \delta\phi_k \delta\phi_{k'} \rangle \,, \tag{11.46}$$

which means that the power spectrum for ζ is

$$\mathcal{P}_\zeta(k) = \frac{H^2}{\dot{\phi}^2}\mathcal{P}_\phi(k) = \left(\frac{H^2}{2\pi\dot{\phi}}\right)^2\Bigg|_{k,\text{horizon exit}} . \tag{11.47}$$

This is the power spectrum that appeared in the CMBR oscillations in Chap. 7, so it can be extracted from experimental data. But during slow-roll, as we saw in Chap. 9,

$$\dot{\phi} \simeq -\frac{V'}{3H}; \quad V \simeq 3M_{\rm Pl}^2 H^2. \tag{11.48}$$

In fact, we can directly relate the ζ perturbation to $\delta\phi$ perturbation and the potential V as

$$\zeta(\vec{x}) = -H\frac{\delta\phi(\vec{x})}{\dot{\phi}} \simeq \frac{3H^2}{V'}\delta\phi(\vec{x}) \simeq \frac{V}{M_{\rm Pl}^2 V'}\delta\phi(\vec{x}) = \frac{1}{\sqrt{2\epsilon}}\delta\phi(\vec{x}). \tag{11.49}$$

But more importantly, we can write now the power spectrum for ζ as a function of only parameters in the potential, as

$$\mathcal{P}_\zeta(k) = \left(\frac{H^2}{2\pi\dot{\phi}}\right)^2\Bigg|_{k,\text{horizon exit}} = \frac{1}{M_{\rm Pl}^2}\left(\frac{V}{M_{\rm Pl}V'}\right)^2\left(\frac{H}{2\pi}\right)^2\Bigg|_{k=aH} , \tag{11.50}$$

where we have used the fact that horizon exit corresponds to $k = aH$, as we saw earlier. Finally, using the fact that $(V'/(VM_{\rm Pl}))^2 = 2\epsilon$, we obtain

$$\mathcal{P}_\zeta(k) = \frac{1}{24\pi^2 M_{\rm Pl}^4}\frac{V}{\epsilon}\Bigg|_{k=aH} = \frac{1}{M_{\rm Pl}^2}\left(\frac{H_{\rm infl}}{2\pi\sqrt{2\epsilon}}\right)^2\Bigg|_{k=aH} . \tag{11.51}$$

From CMBR observations, one can extract the fact that $P_\zeta(k)$ is approximately constant, like we also said in Chap. 7. The first satellite that observed the CMBR fluctuations, the COBE satellite, has measured that (from the measurement of $\delta T/T$, like we said in Chap. 7)

$$P_\zeta(k) \simeq 2.4 \times 10^{-9} \,, \tag{11.52}$$

which means that

$$\frac{1}{M_{\rm Pl}^2}\left(\frac{H_{\rm infl}}{2\pi\sqrt{2\epsilon}}\right)^2\Bigg|_{k=aH} \simeq 2.9\times 10^{-9}. \tag{11.53}$$

This doesn't directly constrain the scale of inflation, but rather only its ratio to ϵ, specifically

$$\left(\frac{V}{\epsilon}\right)^{1/4}\Bigg|_{k=aH} \simeq 6.6\times 10^{16}\,\text{GeV}. \tag{11.54}$$

The power spectrum $\mathcal{P}_\zeta(k)$ is not completely flat however, it has a small *spectral tilt*, defined as

$$n_s - 1 \equiv \frac{d\ln\mathcal{P}_\zeta(k)}{d\ln k}\,, \tag{11.55}$$

such that

$$\mathcal{P}_\zeta(k) \propto k^{n_s-1}. \tag{11.56}$$

From the formula for $\mathcal{P}_\zeta(k)$, and considering that $k = aH$ (the formula is at horizon exit), we find

$$n_s - 1 = 2\frac{d\ln H}{d\ln(aH)} - \frac{d\ln\epsilon}{d\ln(aH)}. \tag{11.57}$$

But since $H \simeq$const.,

$$d(\ln aH) \simeq \frac{da}{a} \simeq Hdt - dN\,, \tag{11.58}$$

where N is the number of e-folds.

Then we find

$$n_s - 1 = -2\frac{d\ln H}{dN} + \frac{d\ln\epsilon}{dN} = -2 - (4\epsilon - 2\tilde{\eta}) = (-6\epsilon + 2\tilde{\eta})|_{k=aH}\,, \tag{11.59}$$

where we have used formulas from Chap. 9 about $d(\ln H/)dN$ and $d\ln\epsilon/dN$.

Since from Planck 2018 data [77]

$$n_s = 0.965 \pm 0.004\,, \tag{11.60}$$

we can impose constraints on ϵ and $\tilde{\eta}$.

Moreover, in a similar way we can now calculate the running of the spectral tilt with scale,

$$\frac{dn}{d\ln k} = -6\frac{d\epsilon}{(-dN)} + 2\frac{d\tilde{\eta}}{(-dN)} = -6\epsilon(4\epsilon - 2\tilde{\eta}) + 2(2\epsilon\tilde{\eta} - \xi^2) = 16\epsilon\tilde{\eta} - 24\epsilon^2 - 2\xi^2. \tag{11.61}$$

11.6 Primordial Tensor Perturbations

Until now we have described the scalar perturbations, which have already been observed in the CMBR. But in principle there could be also tensor perturbations, which could be observed. In fact, the BICEP2 experiment claimed to have seen them, but then in turned out that they had not properly accounted for backgrounds.

Tensor perturbations arise as perturbations in the metric, as we saw in Chap. 7, of the type

$$ds^2 = a^2(\eta)[-d\eta^2 + (\delta_{ij} + h_{ij})dx^i dx^j] \,, \tag{11.62}$$

and by writing the Einstein equations we see that they satisfy also a quadratic equation, of the type

$$\ddot{h}_{ij} + 2aH\dot{h}_{ij} + k^2 h_{ij} = 8\pi G_N T^T_{ij} \,. \tag{11.63}$$

But more importantly, the tensor perurbations correspond to two propagating degrees of freedom (graviton polarizations),

$$h^i{}_j = h_+ e_+{}^i{}_j + h_\times e_\times{}^i{}_j \,. \tag{11.64}$$

In fact, substituting the perturbed ansatz in the Einstein-Hilbert action, we find that the action for the graviton modes $h_+, h_\times$ is just the (rescaled) action for scalar modes, with canonical scalars

$$\phi_{+,\times} = \frac{M_{\rm Pl}}{\sqrt{2}} h_{+,\times} \,. \tag{11.65}$$

But that means that the two-point function for $\phi_{+,\times}$ is the same as the one of ϕ_k, so

$$\langle \phi_+(\vec{k})\phi_+(\vec{k}')\rangle = \langle \phi_\times(\vec{k})\phi_\times(\vec{k}')\rangle = \frac{2\pi^2}{k^3}\left(\frac{H}{2\pi}\right)^2 \delta^3(\vec{k}+\vec{k}') \,. \tag{11.66}$$

On the other hand, the power spectrum for the graviton modes is defined by

$$\langle 2h_+(\vec{k})2h_+(\vec{k}')\rangle = \langle 2h_\times(\vec{k})2h_\times(\vec{k}')\rangle = \frac{2\pi^2}{k^3}\mathcal{P}_h(k)\delta^3(\vec{k}+\vec{k}') \,, \tag{11.67}$$

from which we find the power spectrum for the graviton

$$\mathcal{P}_h(k) = \frac{8}{M_{\rm Pl}^2}\left(\frac{H}{2\pi}\right)^2\Bigg|_{k=aH} \,. \tag{11.68}$$

Note that this power spectrum, unlike the one for ζ_k, has no ϵ in it, so it allows us to directly measure the scale of inflation H_*. The reason for this fact is that the

graviton was proportional to ϕ_k, whereas ζ_k involved a factor of $H/\dot{\phi} = 1/\sqrt{2\epsilon}$ in its relation to $\delta\phi_k$.

The amplitude of the tensor perturbations is usually defined as a ratio to the amplitude of the scalar ones,

$$r = \frac{\mathcal{P}_h}{\mathcal{P}_\zeta}\,, \tag{11.69}$$

so we see that we have

$$r = 16\epsilon. \tag{11.70}$$

We can also define a spectral tilt for the tensor perturbation power spectrum in the same way as it was defined for the scalar perturbations,

$$n_T - 1 = \left.\frac{d\ln\mathcal{P}_h(k)}{d\ln k}\right|_{k=aH} = \frac{2d\ln H}{H dt} = \frac{2\dot{H}}{H^2} \equiv -2\epsilon_H. \tag{11.71}$$

Note that this is the same calculation as for the scalar perturbation, just with only the first term, since there is no epsilon in the graviton power spectrum. During slow-roll, $\epsilon_H \simeq \epsilon$, so

$$n_T - 1 \simeq -2\epsilon. \tag{11.72}$$

Lyth Bound

We have seen previously the relation between the number of e-folds and the variation in the canonical scalar,

$$N = \int_{\phi_i}^{\phi_f} \frac{d\phi/M_{\rm Pl}}{\sqrt{2\epsilon}}. \tag{11.73}$$

If $\epsilon \simeq$ constant, in a given region, then

$$\Delta N \simeq \frac{\Delta\phi}{\sqrt{2\epsilon}M_{\rm Pl}}. \tag{11.74}$$

Specifically, we are interested in the CMBR scales for $1 \leq l \leq 200$ or so, which are among the earliest to get outside the horizon during inflation (first one in, last one out), so that during that time slow-roll is very good, so ϵ is indeed constant to a good approximation.

Then the above relation gives a sort of limit on the total field variation, given an amount of e-folds (which also has a minimum itself),

$$\Delta\phi \geq M_{\rm Pl} \times \Delta N\sqrt{2\epsilon}. \tag{11.75}$$

But since we have just seen that $r = 16\epsilon$, we get that

$$\Delta\phi \geq M_{\rm Pl} \times \sqrt{r} \times \frac{\Delta N}{\sqrt{8}}. \tag{11.76}$$

But that in turn means that if we detect an observable amount of tensor perturbations, let's say $r \sim 0.1$, we need to have a trans-Planckian excursion for the canonical field, in this case

$$\Delta\phi \geq 5M_{\rm pl}. \tag{11.77}$$

That can be problematic, since we expect to generically have uncontrollable quantum gravity corrections in this case.

11.7 Constraints on Inflationary Models

Finally, we make a review of the constraints imposed by experimental data on the inflationary models.

–Slow-roll (which is the standard inflationary paradigm) requires $\epsilon, \tilde{\eta} \ll 1$.

–We need to have sufficient inflation, which amounts to the constraint on the number of e-folds, usually $\Delta N \geq \Delta N_{\rm min} \sim 60$ or so.

–We need to fix the amplitude of the scalar perturbation (the "COBE normalization"), $\mathcal{P}_\zeta(k) \sim 2.4 \times 10^{-9}$.

–We need to match the experimental data for $n_s - 1$, and the bounds on r, implying conditions on ϵ and $\tilde{\eta}$.

–We need to have sufficient reheating, and a reheating temperature certainly larger than the BBN temperature, $T_{RH} \geq 1\,\text{MeV}$, but usually $T_{RH} \geq 10^9\,\text{GeV}$ or so, as we saw.

–We need to get rid of unwanted relics, which means we need to suppress them by a factor of $10^{18} T_{RH}/10^{10}\,\text{GeV}$ during inflation.

–We need to suppress quantum gravity corrections due to the Lyth bound, and in general quantum corrections. We will see that this is related to the so-called η problem. We will say more about this in part III, when discussing string inflation.

–Finally, it would be ideal if the model we have is part of a good particle physics model. (though this is not always imposed by cosmologists, more by string theorists and perhaps by particle theorists).

Important Concepts to Remember

- The scales of the CMBR (about 3000 Mpc) have just got inside the horizon, so they were among the first to exit during inflation (first out, last in).
- Scalar fluctuations during inflation obey an equation that is oscillatory on highly subhorizon scales, and constant on highly superhorizon scales.
- A classical solution that was $\sim e^{-ik\eta}$ on subhorizon scales corresponds to an (initial) quantum state of the Bunch-Davies vacuum $|0\rangle$.
- Thus the initial quantum fluctuations due to the vacuum state get outside the horizon, where they grow and become classical. They return inside the horizon as classical fluctuations, that resume their oscillations.

- The power spectrum of the scalar is flat, $\mathcal{P}_\phi(k) = (H/(2\pi))^2 \simeq$constant.
- In reality, we have a scalar-gravity coupled system, and the common mode satisfies the Mukhanov-Sasaki equation. One can write down an equation for the gauge invariant curvature perturbation ζ_k, the same obtained by replacing $\varphi = a\delta\phi = z\zeta$, with $z \equiv a\dot{\phi}/H$.
- The power spectrum for ζ at horizon exit is $\mathcal{P}_\zeta(k) = (H/(2\pi\dot{\phi}))^2|_{k=aH} = (H/(2\pi\sqrt{2\epsilon}M_{\rm Pl}))^2|_{k=aH}$.
- From the CMBR observations, the power spectrum of ζ is approximately flat, of 2.4×10^{-9}, with a spectral tilt of $n_s - 1 = -6\epsilon + 2\tilde{\eta}$.
- Tensor perturbations correspond to two scalar perturbations for the graviton polarizations.
- One obtains $\mathcal{P}_h = 8(H/(2\pi M_{\rm Pl}))^2$ independent on ϵ, so $r = \mathcal{P}_h/\mathcal{P}_\zeta = 16\epsilon$.
- The Lyth bound says that $\Delta\phi/M_{\rm Pl} \geq \sqrt{r}\delta N/\sqrt{8}$, so a large observed r implies a trans-Planckian excursion for the scalar field, so we have to take care of quantum corrections (including quantum gravity ones).
- On inflationary models, we need slow-roll constraints, the constraint on the number of e-folds, on $n_s - 1$ and r, on T_{RH}, on relics. We need to take care of the Lyth bound (quantum corrections), and perhaps embed in a good particle physics model.

Further reading: See Chaps. 24, 25 in [9], Chap. 12 in [11], Chap. 10 in [12].

Exercises

(1) Consider $V(\phi) = \lambda\phi^p$. Calculate $\epsilon, \eta, n_s, r, N_e$ and $\Delta\phi$ and impose the experimental constraints on the result.
(2) Find the general solution for the fluctuation $\varphi_k(\eta)$ of the massless field, and show that it reduces to the planar wave as $\eta \to -\infty$.
(3) Show that the Mukhanov-Sasaki equation,

$$\frac{d^2\zeta_k}{d\eta^2} + \frac{2}{z}\frac{dz}{d\eta}\frac{d\zeta_k}{d\eta} + k^2\zeta_k = 0 \tag{11.78}$$

has, in the slow-roll regime, the solution

$$\zeta_k = -\frac{\sqrt{\pi}}{2(2\pi)^{3/2}}\frac{\sqrt{\eta}}{z(\eta)}e^{i\pi\frac{\nu}{2}+i\frac{\pi}{4}}H_\nu^{(1)}(-k\eta). \tag{11.79}$$

(4) Consider a cosine potential

$$V = V_0[1 + \cos(a\phi)]. \tag{11.80}$$

Given an observed tensor perturbation of $r = 0.01$, what is the constraint on a such that we have a sub-Planckian scalar field excursion? How about if we sum two of these potentials,

$$V = V_{0,1}[1 + \cos(a_1\phi_1)] + V_{0,2}[1 + \cos(a_2\phi_2)] \,, \tag{11.81}$$

can we improve the constraints on a_1, a_2 for any fine-tuning of the parameters (and any scalar field combination)?

Part II
Elements of String Theory

Chapter 12
Extra Dimensions and Kaluza–Klein

In this chapter we will study the idea of extra dimensions, as proposed by Kaluza and Klein.

The idea of having extra dimensions, besides our usual $3 + 1$, is an old one. Already Theodor Kaluza (in 1921) and Oskar Klein (in 1926) had considered it. The general idea of Klein was to consider that our space is a product of our four dimensional space M_4 and a small compact space K_n, i.e., $K_D = M_4 \times K_n$, with $D = 4 + n$. Then if the volume (or rather, the radii) of the extra dimensional space is really small, they would be unobservable. Perhaps the size is comparable to the Planck length l_{Pl}, which would be natural in a quantum gravity theory, or maybe otherwise larger, yet still small. The compact space is something like a square torus $T^n = S^1 \times \cdots \times S^1$, or maybe a sphere S^n, or some more complicated thing like $\mathbb{CP}^n$, CY_n, etc.

But the reason why it is interesting is the reason that Kaluza proposed the simplest model, with S^1 space (a simple circle), namely that we can can use this Kaluza–Klein (KK) theory for unification of the fields.

In each point in our $3 + 1$ dimensional space, $(\vec{x}, t)$, we have a space K_n, that can perhaps even depend on it, $K_n(\vec{x}, t)$. This will make a general $4 + n$ dimensional spacetime K_D with points $(\vec{x}, t, \vec{y})$, but for KK theory, more precisely we want to have a product spacetime $K_D = M_4 \times K_n$.

Mathematically, we can say that functions live in the total spacetime K_D, instead of M_4, so $\phi(\vec{x}, t, \vec{y})$ instead of $\phi(\vec{x}, t)$.

In the simplest case considered by Kaluza, of $K_n = S^1$ (a circle), we know that there is the Fourier theorem, saying that any function on the total space can be expanded in Fourier modes on the circle,

$$\phi(\vec{x}, t; y) = \sum_{n \geq 0} e^{\frac{iny}{R}} \phi_n(\vec{x}, t). \tag{12.1}$$

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_12

That means that under the Fourier expansion, we can reduce a field in the larger space to an infinite, countable, set of 3+1 dimensional functions $\phi_n(\vec{x}, t)$.

The case of a general compact space K_n instead of the circle S^1 is a generalization of the above case.

12.1 KK Metrics

In the Kaluza–Klein theory, there are 3 metrics that are sometime called KK metrics.

1. **The KK Background Metric**

We need to start with a space describing the product $M_4 \times K_n$, that is a background space. It usually needs to be a solution of the equations of motion (Einstein's equations), though there are cases when one can consider also backgrounds that do not solve the equations of motion. Since the background is of product type, we have

$$g_{\Lambda\Sigma}(\vec{x}, t; \vec{y}) = \begin{pmatrix} g^{(0)}_{\mu\nu}(\vec{x}, t) & \mathbf{0} \\ \mathbf{0} & g^{(0)}_{mn}(\vec{y}) \end{pmatrix}. \tag{12.2}$$

Here $\vec{y}$ are coordinates on K_n, so $g^{(0)}_{mn}(\vec{y})$ is a metric on it.

Note that the metric itself is one of the fields of the theory, so when we write $K_D = M_4 \times K_n$ we mean the background is a product, but the full metric can include fluctuations of all types.

2. **The KK Expansion**

We next consider the full fluctuating solution for the metric and other fields, including the background and the fluctuations around it. Then the KK expansion is simply a generalization of the Fourier theorem for any compact space.

Indeed, after the Fourier case (12.1), the next one is the case of expansion on S^2, known from quantum mechanics (though there we didn't have dependence on $(\vec{x}, t)$). This is the expansion in the spherical harmonics $Y_{lm}(\theta, \phi)$ on S^2,

$$\phi(\vec{x}, t; \theta, \phi) = \sum_{lm} \phi_{lm}(\vec{x}, t) Y_{lm}(\theta, \phi). \tag{12.3}$$

Again this is an identity (it is a theorem that we can always write this expansion). In both of these cases, the functions into which we expand are eigenfunctions of the Laplacean,

$$\begin{aligned} \partial_y^2 e^{\frac{iny}{R}} &= -\left(\frac{n}{R}\right)^2 e^{\frac{iny}{R}} \\ \Delta_2 Y_{lm}(\theta, \phi) &= -\frac{l(l+1)}{R^2} Y_{lm}(\theta, \phi). \end{aligned} \tag{12.4}$$

We moreover note that n and l define the eigenvalue of the Laplacean, as well as the dimension of the $(2l+1$ dimensional) spin l representation of $SO(3) = SU(2)$, the symmetry group of the 2-sphere, whereas m is an index in the representation.

Similarly then, in the general case we write an expansion in "spherical harmonics" $Y_q^{I_q}(\vec{y})$, where q measures the eigenvalue of the Laplacean Δ_n as well as the dimension of the representation, and I_q is an index in the representation of the symmetry group of the compact space. Therefore we write

$$\phi(\vec{x}, t; \vec{y}) = \sum_{q, I_q} \phi_q^{I_q}(\vec{x}, t) Y_q^{I_q}(\vec{y}). \tag{12.5}$$

The Y_q are eigenfunctions of the Laplacean,

$$\Delta_n Y_q^{I_q}(\vec{y}) = -m_q^2 Y_q^{I_q}(\vec{y}). \tag{12.6}$$

From a 4 dimensional point of view, considering a single mode in the expansion of the field,

$$\phi(\vec{x}, t; \vec{y}) = \phi_q^{I_q}(\vec{x}, t) Y_q^{I_q}(\vec{y}) \ , \tag{12.7}$$

if we act with the D-dimensional D'Alembertian $\Box_D = \Box_4 + \Delta_n$ (since the space is a product, the D'Alembertian splits into the sum of a 4 dimensional D'Alembertian and an n-dimensional Laplacean), we find

$$\Box_D \phi(\vec{x}, t; \vec{y}) = (\Box_4 + \Delta_n)\phi(\vec{x}, t; \vec{y}) = (\Box_4 - m_q^2)\phi(\vec{x}, t; \vec{y}). \tag{12.8}$$

That means that if the field is D-dimensional massless,

$$\Box_D \phi(\vec{x}, t; \vec{y}) = 0 \ , \tag{12.9}$$

then the 4-dimensional field is massive with mass m_q,

$$[(\Box_4 - m_q^2)\phi_q^{I_q}(\vec{x}, t)] Y_q^{I_q}(\vec{y}). \tag{12.10}$$

That means that if we want to see some structure in K_n, at the level of the field $\phi_q^{I_q}$, or rather of the spherical harmonic $Y_q^{I_q}$, we need to have at least an energy of m_q to see it.

3. **The KK Reduction Ansatz**

Which brings us to the KK reduction ansatz. If we are at energies much smaller than *all* the m_q, which as we saw are $\propto 1/R$ (the scale of the compact space), we will only excite the massless mode $\phi_0^{I_0} Y_0^{I_0}$, and we will not be able to see structure in the compact space K_n, i.e., we are not able to probe it.

In effect, we have *dimensionally reduced* the D-dimensional theory to 4 dimensions. That means that we could put all the nonzero fields to zero and only keep the

"$n = 0$" representation (multiplet of fields). This is maybe "independent on $\vec{y}$", or at least has a spherical harmonic with the simplest possible dependence on $\vec{y}$, $Y_0(\vec{y})$.

Therefore the dimensional reduction ansatz is

$$\phi(\vec{x}, t; \vec{y}) = \phi_0(\vec{x}, t) Y_0(\vec{y}). \tag{12.11}$$

But note that this is an *ansatz*, which is German for "guess", i.e., it is not guaranteed to work, we need to check whether it is consistent. We will have more to say on that later.

In conclusion, we can say that: the KK background is a solution; the KK expansion is a parametrization; and the KK reduction is an ansatz.

12.2 Fields with Spin

Until now we have implicitly assumed that the fields were scalars under Lorentz transformations, but the original motivation for KK theory was to unify fields with spin into a single field with spin in higher dimensions, so we need to consider them now. Unification occurs because various components of the higher dimensional field act as different fields in 4 dimensions.

The simplest example is the one of electromagnetism with vector field A_M, and $M = (\mu, m)$, where $\mu = 0, 1, 2, 3$ and $m = 4, \ldots, 4 + n$. Then the gauge (vector) field A_M splits as

$$A_M(\vec{x}, t; \vec{y}) = (A_\mu(\vec{x}, t; \vec{y}), A_m(\vec{x}, t; \vec{y})). \tag{12.12}$$

Here A_μ is a 4 dimensional gauge (vector) field, since under an element $\Lambda_\mu{}^\nu \in SO(1, 3)$ it transforms as a vector

$$A'_\mu(\vec{x}', t'; \vec{y}) = \Lambda_\mu{}^\nu A_\nu(\vec{x}, t; \vec{y}). \tag{12.13}$$

(this happens since A_M transforms as a vector under $\Lambda_M{}^N \in SO(1, 3 + n)$, and $SO(1, 3) \subset SO(1, 3 + n)$). On the other hand, A_m is a 4 dimensional scalar, since it is invariant under the same,

$$A'_m(\vec{x}', t'; \vec{y}) = A_m(\vec{x}, t; \vec{y}). \tag{12.14}$$

Indeed, we assume that the theory in D dimensions is Lorentz invariant, i.e., under $SO(1, 3 + n)$, and that the KK expansion (and certainly the KK dimensional reduction ansatz) breaks the Lorentz group to $SO(1, 3) \times SO(n)$ (as does the KK background metric). Here $SO(n)$ only acts on A_m, as an *internal* symmetry from the point of view of $3 + 1$ dimensions,

$$A'_m(\vec{x}, t; \vec{y}') = \Lambda_m{}^n A_m(\vec{x}, t; \vec{y}). \tag{12.15}$$

This is one of the important characteristics of the KK program: the geometrization of internal symmetries. Indeed, in the Standard Model we have internal (non-spacetime) symmetries, like the local $SU(3) \times SU(2) \times U(1)$, but now these internal symmetries appear from the symmetries of the compact space (Lorentz symmetries, or isometries).

We can write the the KK expansion for the components of A_M in the usual way,

$$\begin{aligned} A_\mu(\vec{x}, t; \vec{y}) &= \sum_{q, I_q} A_\mu^{q, I_q}(\vec{x}, t) Y_q^{I_q}(\vec{y}) \\ A_m(\vec{x}, t; \vec{y}) &= \sum_{q, I_q} A^{q, I_q}(\vec{x}, t) Y_m^{q, I_q}(\vec{y}). \end{aligned} \tag{12.16}$$

Note that the internal Lorentz symmetry $SO(n)$ acts on $Y_m^{q I_q}$ only. m is an index in the (local) Lorentz symmetry of the space, the symmetry of the space tangent to K_n at that point, whereas I_q is an index in the *isometry* of the space. For an n-sphere $S^n = SO(n+1)/SO(n)$, the isometry group is $SO(n+1)$, but the local Lorentz group is $SO(n)$, so m would be in $SO(n)$, but I_q would be in $SO(n+1)$. The two groups only coincide in the simplest case of the torus T^n, when both groups are $SO(n)$, and $Y_q^{I_q} \propto$ "$\delta_m^{I_q}$". We then find that A^{q, I_q} are 4 dimensional scalars in the representation I_q of the isometry group, and A_μ^{q, I_q} are vectors in the representation of the isometry group.

The next interesting case is of gravity (symmetric traceless tensor for the metric fluctuation). Again different components have different 4 dimensional spins, and the decomposition is

$$g_{MN}(\vec{x}, t; \vec{y}) = \begin{pmatrix} g_{\mu\nu}(\vec{x}, t) & g_{\mu m}(\vec{x}, t; \vec{y}) \\ g_{m\mu}(\vec{x}, t; \vec{y}) & g_{mn}(\vec{x}, t; \vec{y}) \end{pmatrix}. \tag{12.17}$$

Then the fluctuation in $g_{\mu\nu}$ is a $SO(1,3)$ (4 dimensional Lorentz) symmetric traceless tensor, i.e., a graviton (so the full variable is the 4 dimensional metric), $g_{\mu m} = g_{m\mu}$ is a $SO(1,3)$ vector, i.e., gauge field, transforming as

$$g'_{\mu m}(\vec{x}', t'; \vec{y}) = \Lambda_\mu{}^\nu g_{\nu m}(\vec{x}, t; \vec{y}) \ , \tag{12.18}$$

and $g_{mn}(\vec{x}, t; \vec{y})$ are $SO(1,3)$ (4 dimensional Lorentz) scalars. As before, the KK expansion (and the background metric, really) break the local Lorentz symmetry in D dimensions, $SO(1, 3+n)$, to $SO(1,3) \times SO(n)$, 4 dimensional Lorentz times internal symmetry. Here $SO(n)$ acts on $\vec{y}$ only, through the m, n indices on the spherical harmonics, for instance $Y_{mn}^{q, I_q}(\vec{y})$ for the scalars.

12.3 The Original Kaluza–Klein Theory

The idea of Kaluza was to unify gravity ($g_{\mu\nu}$) and electromagnetism (B_μ) in a 5 dimensional metric (gravity) $g_{\Lambda\Sigma}$.

1. Writing the KK background metric for $M_4 = Mink_4 \times S^1$, we have

$$g^{(0)}_{\Lambda\Sigma}(\vec{x}, t; \vec{y}) = \begin{pmatrix} \eta_{\mu\nu} & \mathbf{0} \\ \mathbf{0} & g_{55} = 1 \end{pmatrix}. \tag{12.19}$$

2. The KK expansion (parametrization) is now simply the Fourier expansion on the circle,

$$g_{\Lambda\Sigma}(\vec{x}, t; y) = \begin{pmatrix} \eta_{\mu\nu} + \sum_{n\geq 0} h^{(n)}_{\mu\nu}(\vec{x}, t) e^{\frac{iny}{R}} & g_{\mu 5} \\ g_{\mu 5}(\vec{x}, t; y) = \sum_{n\geq 0} h^{(n)}_{\mu 5}(\vec{x}, t) e^{\frac{iny}{R}} & g_{55} = 1 + \sum_{n\geq 0} \phi^{(n)}(\vec{x}, t) e^{\frac{iny}{R}} \end{pmatrix}. \tag{12.20}$$

3. The KK reduction ansatz amounts to keeping only the $n = 0$ mode, so

$$g_{\Lambda\Sigma}(\vec{x}, t; y) = \begin{pmatrix} \eta_{\mu\nu} + h^{(0)}_{\mu\nu}(\vec{x}, t) \equiv g_{\mu\nu}(\vec{x}, t) & g_{\mu 5} \\ g_{\mu 5}(\vec{x}, t; y) = h^{(0)}_{\mu 5}(\vec{x}, t) \equiv B_\mu & g_{55} = 1 + \phi^{(0)}(\vec{x}, t) \equiv \varphi(\vec{x}, t) \end{pmatrix}. \tag{12.21}$$

But we see that we wanted to unify gravity and electromagnetism, and we seem to need to have also a scalar. Yet there is no observed scalar, until now, except the Higgs, which is extremely massive compared to the nearly massless scalar needed above. And moreover, the existence of a nearly massless scalar would mean a fifth force, which certainly contradicts experiment. So one needs to get rid of this potential scalar.

The solution proposed originally was simply to set $\varphi = 1$ ($\phi^{(0)} = 0$). However, while the original KK dimensional reduction is actually consistent, i.e., it solves the equations of motion, this further reduction is not consistent. We left the check of this statement (at the linearized level) as an exercise. We will say more on consistent truncations in the next chapter.

12.4 Generalization: The n-Torus $T^n = (S^1)^{\times n}$

The simplest generalization is obtained by considering instead of a circle, a product of circles, i.e., the square torus. Indeed, a general torus is obtained by periodic identifications. In particular, for the square torus, we have the identifications $y_i \sim y_i + 2\pi R_i$ (in general, we can have identifications at angles). The bottom line is that the torus is flat, $g^{(0)}_{mn} = \delta_{mn}$, since we have a piece of $\mathbb{R}^n$.

The *KK background metric* is then

$$g^{(0)}_{\Lambda\Sigma}(\vec{x}, t; \vec{y}) = \begin{pmatrix} g^{(0)}_{\mu\nu}(\vec{x}, t) & \mathbf{0} \\ \mathbf{0} & \delta_{mn} \end{pmatrix}. \tag{12.22}$$

The *KK expansion* is the product of Fourier expansions on the circles of radii R_i in T^n, i.e.,

$$g_{\Lambda,\Sigma}(\vec{x}, t; \vec{y}) = \begin{pmatrix} g^{(0)}_{\mu\nu}(\vec{x}, t) + \sum_{\{n_i\}} h^{\{n_i\}}_{\mu\nu}(\vec{x}, t) \prod_{i=1}^n e^{\frac{i n_i y_i}{R_i}}; & g_{\mu m}(\vec{x}, t; \vec{y}) \\ g_{\mu m}(\vec{x}, t; \vec{y}) = \sum_{\{n_i\}} B^{m,\{n_i\}}_{\mu}(\vec{x}, t) \prod_{i=1}^n e^{\frac{i n_i y_i}{R_i}}; & g_{mn} = \delta_{mn} + \sum_{\{n_i\}} h^{\{n_i\}}_{mn}(\vec{x}, t) \prod_{i=1}^n e^{\frac{i n_i y_i}{R_i}} \end{pmatrix}. \tag{12.23}$$

We see that here the spherical harmonics are

$$Y_{\{n_i\}}(\vec{y}) = \prod_{i=1}^n e^{\frac{i n_i y_i}{R_i}}. \tag{12.24}$$

The *KK reduction ansatz* amounts to taking just the zero modes, so

$$g_{\Lambda\Sigma}(\vec{x}, t; \vec{y}) = \begin{pmatrix} g^{(0)}_{\mu\nu} + h^{\{0\}}_{\mu\nu}(\vec{x}, t) \equiv g_{\mu\nu}(\vec{x}, t); & g_{\mu m}(\vec{x}, t) \\ g_{\mu m}(\vec{x}, t) = B^{m\{0\}}_{\mu}(\vec{x}, t); & g_{mn}(\vec{x}, t) = \delta_{mn} + h^{\{0\}}_{mn}(\vec{x}, t) \end{pmatrix}. \tag{12.25}$$

Then $g_{\mu\nu}(\vec{x}, t)$ is the 4 dimensional metric, $B^{m\{0\}}_{\mu}(\vec{x}, t)$ are vectors and $h^{\{0\}}_{mn}(\vec{x}, t)$ are scalars.

Important Concepts to Remember

- In Kaluza–Klein theory, the space is a product of a 4 dimensional space and a small n-dimensional compact space, $K_D = M_4 \times K_n$.
- We don't see the extra dimensions because they are small (small radii), and the fields are unified, in the sense that various components of a field with spin in the higher dimension act as various fields with spin in 4 dimensions.
- The KK theory starts with the KK background metric, usually a solution to the equations of motion (Einstein's equations), of the product form.
- We then do a KK expansion around it, in spherical harmonics, which is a generalization of the Fourier expansion. It is therefore a theorem, something that we can always do on a compact space, or another words a parametrization.
- The spherical harmonics $Y^{I_q}_q(\vec{y})$ are eigenfunctions of the Laplacean, where q defines the eigenvalue of the Laplacean, and I_q is an index in a representation of the symmetry group of K_n.
- A D-dimensional massless field is 4-dimensional massive with m_q, the eigenvalue of the Laplacean being $-m_q^2$.

- The KK reduction ansatz is an ansatz ("guess"), saying that we can truncate the series only to the first term, the "$n = 0$" representation, with "y-independent" spherical harmonic, or rather with the simplest $Y_0(\vec{y})$.
- The split of fields with spin, where the various components of the D dimensional field become various fields with spin in 4 dimensions, breaks the local Lorentz symmetry as $SO(1, 3 + n) \to SO(1, 3) \times SO(n)$.
- The $SO(n)$ symmetry appears as an internal symmetry, thus the KK theory geometrizes internal symmetries.
- The indices m in $SO(n)$ appear on the spherical harmonics, like for the scalars in g_{mn} we have $Y^{q,I_q}_{mn}(\vec{y})$, whereas I_q are indices in the isometry group of K_n. For the torus, the two are identified.
- In the original KK theory, for dimensional reduction, besides $g_{\mu\nu}$ and B_μ (gravity and electromagnetism), we get a massless scalar, that cannot be put to zero (it is inconsistent to do so).
- For a torus, the spherical harmonics are the products of the Fourier modes.

Further reading: See [13] for the KK approach to supergravity, and a quick review of the KK program.

Exercises

(1) Write the dimensional reduction on S^1 of a 5 dimensional field satisfying "self-duality in odd dimensions", with action

$$S = \int d^5x \left[\epsilon^{\mu\nu\rho\sigma\tau} S_{\mu\nu}\partial_\rho S_{\sigma\tau} + \sqrt{-g}m^2 S_{\mu\nu}S^{\mu\nu}\right]. \tag{12.26}$$

(2) Write down the KK background metric and KK expansion for a compactification of a 10 dimensional metric on $M^4 \times S^2 \times S^2 \times T^2$.
(3) Show explicitly that putting $\phi = 1$ in the original KK reduction (to linearized level) is inconsistent.
(4) Dimensionally reduce the 11 dimensional 3-form gauge field A_{MNP}, with action

$$\int d^{11}x \left[-\frac{1}{2\cdot 4!}F_{MNPQ}F^{MNPQ} - \frac{1}{3}\epsilon^{M_1\ldots M_{11}}A_{M_1M_2M_3}A_{M_4\ldots M_7}A_{M_8\ldots M_{11}}\right], \tag{12.27}$$

where $F_{MNPQ} = 4\partial_{[M}A_{NPQ]}$, on a torus T^7, down to 4 dimensions.
(5) Prove that if we split the off-diagonal metric on a compact space as $g_{\mu m}(x, y) = B^{AB}_\mu(x)V^{AB}_m(y)$, where V^{AB}_m are Killing vectors of the compact space satisfying $D_nV_m + D_nV_m = 0$, then the general coordinate transformation with $\xi_m(x, y) = \lambda^{AB}(x)V^{AB}_m(y)$ gives a nonabelian gauge transformation of B^{AB}_μ, with parameter $\lambda^{AB}(x)$ [Hint: use the fact that $V^{AB} = V^{mAB}\partial_m$ satisfies the nonabelian algebra].

Chapter 13
Electromagnetism and Gravity in Various Dimensions. Consistent Truncations

In this chapter, we will first generalize electromagnetism and gravity to general dimensions, and then we will define consistent truncations of field theories, with emphasis on the KK reduction of supergravity theories.

13.1 Review of Electromagnetism and Gravity

Electromagnetism

The Maxwell's equations, for the equations of motion

$$\partial^\mu F_{\mu\nu} = j_\nu \ , \tag{13.1}$$

and the Bianchi identities,

$$\partial_{[\mu} F_{\nu\rho]} = 0 \ , \tag{13.2}$$

solved by the existence of a gauge field A_μ,

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \ , \tag{13.3}$$

are valid in any number of dimensions, as we can see (they do not refer to dimensionality anywhere).

For 3 spatial dimensions, we define the electric field,

$$E^i = F^{0i} = -F_{0i} \ , \tag{13.4}$$

and in terms of the gauge field

$$\vec{E} = -\partial_0 \vec{A} - \vec{\nabla}\Phi \ , \tag{13.5}$$

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_13

where $A_\mu = (-\Phi, \vec{A})$. These equations again are independent of dimension, so are valid generally. However, the definition of the magnetic field,

$$B^i = \frac{1}{2}\epsilon^{ijk}F_{jk} \ , \tag{13.6}$$

depends on dimensionality, so is not obviously generalizable. We can of course write a similar relation with the epsilon tensor in other dimensions, but then for instance in 2 spatial dimensions B is a scalar and in 4 spatial dimensions B^{ij} is a tensor.

The local form of Gauss's law is

$$\vec{\nabla}\cdot\vec{E} = \frac{\rho_e}{\epsilon} \Rightarrow \vec{\nabla}^2\Phi = -\frac{\rho_e}{\epsilon}. \tag{13.7}$$

Gravity

The Einstein's equations are

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = 8\pi G_N T_{\mu\nu} \ , \tag{13.8}$$

and are generalized to arbitrary dimension as long as we define $G_N^{(D)}$ in a general dimension D (we see later).

The Riemann tensor in terms of the Christoffel symbol,

$$R^\mu{}_{\nu\rho\sigma} = \partial_\rho\Gamma^\mu{}_{\nu\sigma} - \partial_\sigma\Gamma^\mu{}_{\nu\rho} + \Gamma^\mu{}_{\lambda\rho}\Gamma^\lambda{}_{\nu\sigma} - \Gamma^\mu{}_{\lambda\sigma}\Gamma^\lambda{}_{\nu\rho} \ , \tag{13.9}$$

is also valid in general dimension, as is the Bianchi identity that encapsulates the fact that the Riemann tensor is written in terms of the Christoffel,

$$\partial_{[\lambda}(R^\mu{}_\nu)_{\rho\sigma]} = 0. \tag{13.10}$$

In the Newtonian limit of Einstein gravity, we have the Gauss's law for gravity,

$$\vec{\nabla}^2 U_{\text{Newton}} = 4\pi G_N \rho_m \ , \tag{13.11}$$

where the Newtonian potential is related to the perturbation of the metric by $g_{00} - 1 = \kappa_N h_{00} = -2U_{\text{Newton}}$.

13.2 Interlude: Spheres in Higher Dimensions

Before we continue, we calculate the volume of spheres in arbitrary dimensions.

A sphere in d spatial dimensions (spacetime dimension $D = d + 1$) is defined by

$$S^{d-1}(R) : \sum_{i=1}^{d} x_i^2 = R^2 \ , \tag{13.12}$$

whereas the ball is

$$B^{d-1}(R) : \sum_{i=1}^{d} x_i^2 \leq R^2. \tag{13.13}$$

The volume of the sphere is defined in terms of the volume of the sphere of unit radius as

$$vol(S^{d-1}(R)) = R^{d-1} vol(S^{d-1}). \tag{13.14}$$

For instance, $vol(S^1(R)) = 2\pi R$, $vol(S^2(R)) = 4\pi R^2$.

To find the general case for $vol(S^{d-1})$, we evaluate the d-dimensional Gaussian integral I_d in two different ways. On one hand, the integral is a product of the Gaussian integrals on each direction,

$$I_d = \int_{\mathbb{R}^d} dx_1 \dots dx_d e^{-r^2} = \prod_{i=1}^{d} \int_{-\infty}^{+\infty} dx_i e^{-x_i^2} = (\sqrt{\pi})^d = \pi^{d/2} \ , \tag{13.15}$$

where $r^2 = x_1^2 + \cdots x_d^2$. On the other hand, we can use spherical coordinates and write

$$\begin{aligned} I_d &= \int_0^\infty dr \ vol(S^{d-1}(r)) e^{-r^2} = vol(S^{d-1}) \int_0^\infty dr r^{d-1} e^{-r^2} \\ &= vol(S^{d-1}) \int_0^\infty \frac{dt}{2} e^{-t} t^{\frac{d}{2}-1} = vol(S^{d-1}) \frac{\Gamma(d/2)}{2} \ , \end{aligned} \tag{13.16}$$

where we wrote $r^2 = t$. Equating the two ways of writing I_d, we find

$$vol(S^{d-1}) = \frac{2\pi^{d/2}}{\Gamma(d/2)}. \tag{13.17}$$

We can check that indeed, for $d = 1$, $vol(S^1) = 2\pi/\Gamma(1) = 2\pi$, and for $d = 2$, $vol(S^2) = 2\pi^{3/2}/\Gamma(3/2) = 4\pi$, since $\Gamma(1/2) = \sqrt{\pi}$ and $\Gamma(3/2) = 1/2\Gamma(1/2)$.

13.3 Gauss's Law for Electromagnetism in d Dimensions

3 Spatial Dimensions

The local form of Gauss's law is $\vec{\nabla} \cdot \vec{E} = \rho_e/\epsilon_0$. Integrating it over a ball $B^3(R)$, we find

$$\int_{B^3} dV \ \vec{\nabla} \cdot \vec{E} = \int_{B^3} dV \ \frac{\rho_e}{\epsilon} = \frac{Q}{\epsilon_0}. \tag{13.18}$$

On the other hand, using Stokes's theorem, the left hand side becomes

$$\int_{S^2(R)} d\vec{S} \cdot \vec{E} = vol(S^2) R^2 E(R) \ , \tag{13.19}$$

where because of spherical symmetry, both the area element $d\vec{S}$ and the electric field $\vec{E}$ point radially outwards, hence their dot product equals the normal product. Finally, we find

$$E(R) = \frac{Q}{vol(S^2)\epsilon_0 R^2} = \frac{Q}{4\pi\epsilon_0 R^2}. \tag{13.20}$$

d Spatial Dimensions

Generalizing the above to d spatial dimensions, the local form of Gauss's law is the same, so we integrate it over $B^d(R)$, obtaining

$$\int dV \ \vec{\nabla} \cdot \vec{E} = \int_{B^d} dV \frac{\rho}{\epsilon_0} = \frac{Q}{\epsilon_0}. \tag{13.21}$$

Using Stokes's theorem, the left hand side becomes

$$\int_{\partial B^d(R) = S^{d-1}(R)} d\vec{S} \cdot \vec{E} = vol(S^{d-1}) R^{d-1} E(R) \ , \tag{13.22}$$

so by equating the two sides we get

$$E(R) = \frac{1}{vol(S^{d-1})\epsilon_0} = \frac{\Gamma(d/2)}{2\pi^{d/2}\epsilon_0} \frac{Q}{R^{d-1}}. \tag{13.23}$$

13.4 Gauss's Law for Gravity in d Dimensions

3 Spatial Dimensions

The local form of Gauss's law, integrated over the ball $B^3(R)$, gives

$$\int_{B^3(R)} dV \ \vec{\nabla}^2 U_N = 4\pi G_N \int_{B^3(R)} dV \ \rho_m = 4\pi G_N M. \tag{13.24}$$

Using Stokes's theorem, the left hand side becomes

$$\int_{S^2(R)} d\vec{S} \cdot \vec{\nabla} U_N = vol(S^2) R^2 |\vec{g}(R)| \ , \tag{13.25}$$

where we have defined the gravitational acceleration $\vec{g} = -\vec{\nabla} U_N$. Equating the two sides, we find

$$|\vec{g}(R)| = \frac{4\pi G_N M}{vol(S^2)R^2} = \frac{G_N M}{R^2}. \tag{13.26}$$

Newton's Constant and Planck Constant in Higher Dimensions

Before we move onto higher dimensions, we need to define $G_N^{(D)}$, since only then we can have the same local form of Gauss's law for gravity. It is obtained by writing the same coefficient for the Einstein–Hilbert action in higher dimensions as in 4 dimensions,

$$S_{EH} = \frac{1}{16\pi G_N^{(D)}} \int d^D x \sqrt{-g^{(D)}} R^{(D)}. \tag{13.27}$$

But we need to KK reduce to 4 dimensions, $D = 4 + n$, with $K_D = M_4 \times K_n$,

$$g_{MN} = \begin{pmatrix} g_{\mu\nu}^{(0)} & \mathbf{0} \\ \mathbf{0} & g_{mn}^{(0)} \end{pmatrix}. \tag{13.28}$$

Then we have $\sqrt{-g^{(D)}} = \sqrt{-g^{(4)}}\sqrt{g^{(n)}}$, so by integrating over $d^n x$, the coefficient of $\int d^4x \sqrt{-g^{(4)}} R^{(D)}$ is defined as the 4 dimensional Newton's constant,

$$\frac{1}{G_N^{(D)}} \int d^n x \sqrt{g^{(n)}} = \frac{V^{(n)}}{G_N^{(D)}} \equiv \frac{1}{G_N^{(4)}}\ , \tag{13.29}$$

leading to

$$G_N^{(D)} = G_N^{(4)} V^{(n)}. \tag{13.30}$$

As we have already mentioned, the 4 dimensional Planck scale is constructed out of c, $\hbar$ and $G_N^{(4)}$ as $m_P = \sqrt{\hbar c / G_N^{(4)}}$, and the Planck length is $l_P = \sqrt{\hbar G_N^{(4)}/c^3}$ (so that if $\hbar = c = 1$, $m_P = (G_N^{(4)})^{-1/2}$ and $l_P = (G_N^{(4)})^{1/2}$), but in gravity and cosmology we mostly use the reduced Planck mass, $M_{\rm Pl}^{(4)} = m_P/\sqrt{8\pi}$, so that the coefficient of the Einstein–Hilbert action is $M_{\rm Pl}^2/2$.

In higher dimensions, the coefficient of the Einstein–Hilbert action has dimension $D - 2$, so that the action is defined as

$$\frac{[M_{\rm Pl}^{(D)}]^{D-2}}{2} \int d^D x \sqrt{-g^{(D)}} R^{(D)}. \tag{13.31}$$

Then finally

$$\frac{1}{8\pi G_N^{(D)}} = [M_{\rm Pl}^{(D)}]^{D-2}. \tag{13.32}$$

d Spatial Dimensions

With $G_N^{(D)}$ defined, the local form of the Gauss's law for gravity is the same. Integrating it over the spatial ball $B^d(R)$, we find

$$\int_{B^d(R)} dV\ \vec{\nabla}^2 U_N = 4\pi G_N^{(D)} \int dV\ \rho_m = 4\pi G_N^{(D)} M. \tag{13.33}$$

Using Stokes's theorem, the left hand side becomes

$$\int_{\partial B^d = S^{d-1}(R)} d\vec{S} \cdot \vec{\nabla} U_N = vol(S^{d-1}) R^{d-1} |\vec{g}(R)|. \tag{13.34}$$

Equating the two sides, we obtain

$$|\vec{g}(R)| = \frac{4\pi G_N^{(D)} \Gamma(d/2)}{2\pi^{d/2}} \frac{M}{R^{d-1}} = \frac{\Gamma\left(\frac{D-1}{2}\right)}{2[M_{\rm Pl}^{(D)}]^{D-2} 2\pi^{\frac{D-2}{2}}} \frac{M}{R^{D-2}}. \tag{13.35}$$

Integrating to get the Newton potential, we obtain

$$U_N(R) = -\frac{2\pi^{\frac{3-D}{2}} \Gamma\left(\frac{D-1}{2}\right)}{D-3} \frac{M G_N^{(D)}}{R^{D-3}} = -\frac{\Gamma\left(\frac{D-1}{2}\right)}{4\pi^{\frac{D-1}{2}} (D-3)} \frac{M}{[M_{\rm Pl}^{(D)}]^{D-2} R^{D-3}}. \tag{13.36}$$

For KK compactification on a square torus $T^n = (S_R^1)^{\times n}$, we find

$$G_N^{(D)} = (2\pi R)^n G_N^{(4)}. \tag{13.37}$$

Then the Gauss's law for distances much smaller than the radii, $r \ll R$, is approximately the D-dimensional law,

$$|\vec{g}(r)| = \frac{4\pi \Gamma\left(\frac{D-1}{2}\right)}{2\pi^{\frac{D-1}{2}}} \frac{G_N^{(D)} M}{r^{D-2}}. \tag{13.38}$$

For distances much larger than the radius on the other hand, we have the 4 dimensional law,

$$|\vec{g}(r)| = \frac{G_N^{(4)} M}{r^2} = \frac{G_N^{(D)} M}{(2\pi R)^{D-4} r^2}. \tag{13.39}$$

So the difference between the two is the replacement of r^{D-2} with $R^{D-4} r^2$, up to some numerical factors.

13.5 Consistent Truncations

An issue that arises when considering KK dimensional reduction is the issue of whether the truncation is consistent. Indeed, we need to put all fields with $n > 0$ to zero, but this is not guaranteed to satisfy the original equations of motion, rather it is an ansatz.

When working at the level of the action, what can go wrong is to have terms that are linear in (all the) fields $\phi_{n>0}$ (note that a term with $\phi_1\phi_2$ is quadratic in ϕ_n), which means that when we write the equations of motion for the ϕ_n, we find

$$(\Box - m^2 + \cdots)\phi_n = (\ldots) \ , \tag{13.40}$$

where the left hand side comes from terms quadratic or higher in ϕ_n's, and the right hand side is finite when you put $\phi_n = 0$, so it is inconsistent to put $\phi_n = 0$.

To see an example, consider a ϕ^3 interaction. The KK expansion, substituted in the term $\int d^D x\sqrt{-g^{(D)}}\lambda\phi^3$, leads to terms of the type (we consider the reduction to d dimensions instead of 4, for generality)

$$\lambda \int d^d x\sqrt{-\det g^{(0)}_{\mu\nu}}\phi_q^{I_q}(\vec{x},t)\phi_0^{I_0}(\vec{x},t)\phi_0^{J_0}(\vec{x},t) \times \int d^n x\sqrt{-\det g^{(0)}_{mn}}Y_q^{I_q}(\vec{y})Y_0^{I_0}(\vec{y})Y_0^{J_0}(\vec{y}). \tag{13.41}$$

If this term is nonzero, then the equation of motion of $\phi_q^{I_q}$ will have a term

$$(\Box - m_q^2)\phi_q^{I_q}(\vec{x},t) = \lambda C\phi_0^{I_0}(\vec{x},t)\phi_0^{J_0}(\vec{x},t) \ , \tag{13.42}$$

where

$$C = \int d^n x\sqrt{-\det g^{(0)}_{mn}}Y_q^{I_q}(\vec{y})Y_0^{I_0}(\vec{y})Y_0^{J_0}(\vec{y}). \tag{13.43}$$

Then $\phi_q^{I_q} = 0$ gives a contradiction, since the left-hand side is zero, but the right-hand side is not.

But if $C = 0$, i.e., if the integral of one of the Y_n's with the rest Y_0's vanishes, then the truncation is consistent. This is what happens for instance in the case of the torus T^n, when $Y_0(\vec{y}) = 1$, but then for S^1, $C_1 = \int dy Y^{I_n} = \int dy e^{\frac{iny}{R}} = 0$, and in the T^n case $C_{T^n} = \prod_n C_1 = 0$ again. Thus for the torus there is always a consistent truncation if we keep *all* the $n = 0$ fields. For instance, in the original Kaluza–Klein reduction, $\varphi = 1$ is inconsistent, since it puts one of the $n = 0$ fields to zero.

More generally, if we have invariance under a group G, and we keep *all* the singlets under G in the reduction, the reduction is consistent, since then all

$$\int Y_n Y_0 Y_0 = 0 \ , \tag{13.44}$$

as the Y_n correspond by our assumption to a non-singlet, whereas the Y_0's are singlets. That means that $Y_n Y_0 Y_0$ is a non-singlet, so its integral should be also, except it is a number (which doesn't transform), which means that it must be zero.

Nonlinear Ansatz

Sometimes, when the truncation of the linearized KK expansion is inconsistent, we can write a nonlinear redefinition of the fields that makes it consistent, generically of the type

$$\begin{aligned}\phi'_q &= \phi_q + a\phi_0^2 + \cdots \\ \phi'_0 &= \phi_0 + \sum_{pq,\text{incl.}0} c_{pq}\phi_p\phi_q + \cdots ,\end{aligned} \tag{13.45}$$

where we only wrote the quadratic piece of the redefinition, but there are in principle higher order terms. Or equivalently, we can write from the beginning a *nonlinear KK reduction ansatz*. Note that the linearized KK expansion is always valid (it is a theorem that it is so), but the linearized KK dimensional reduction (the truncation of this expansion to the first terms) is not guaranteed to be correct. A nonlinear ansatz for this KK dimensional reduction instead of the linearized one can sometimes work though.

The simplest example of a nonlinear ansatz is the one needed to get the correct d-dimensional Einstein–Hilbert action for the D-dimensional one. The ansatz is

$$g_{\mu\nu}(\vec{x}, t; \vec{y}) = g_{\mu\nu}(\vec{x}, t) \left[\frac{\det g_{mn}(\vec{x}, t; \vec{y})}{\det g^{(D)}_{mn}(\vec{y})} \right]^{-\frac{1}{d-2}} . \tag{13.46}$$

To fully check this result is somewhat complicated, but we can check it for the case that the square bracket is constant.

Indeed, under a constant rescaling $g_{\mu\nu} \to \lambda g_{\mu\nu}$, the Christoffel symbols $\Gamma \sim g^{-1}\partial g$ and the Ricci scalar $R_{\mu\nu} = R^{\lambda}{}_{\mu\lambda\nu} \sim \partial\Gamma + \Gamma\Gamma$ are invariant. In the Einstein–Hilbert action,

$$R^{(D)} = g^{\Lambda\Sigma} R^{(D)}_{\Lambda\Sigma} = g^{\mu\nu} R^{(D)}_{\mu\nu} + \cdots \tag{13.47}$$

contains the d-dimensional Ricci scalar (plus other terms). Then if we only rescale $g_{\mu\nu}$, but not g_{mn},

$$\sqrt{-g^{(D)}}\, g^{\mu\nu} R^{(D)}_{\mu\nu} = \sqrt{-\det g^{(d)}_{\mu\nu}}\sqrt{\det g_{mn}}\, g^{\mu\nu} R^{(D)}_{\mu\nu} \to \lambda^{\frac{d}{2}-1}\sqrt{-\det g^{(d)}_{\mu\nu}}\sqrt{\det g_{mn}}\, g^{\mu\nu} R^{(D)}_{\mu\nu} , \tag{13.48}$$

so, in order to replace the $\sqrt{\det g_{mn}}$ factor with $\sqrt{g^{(0)}_{mn}}$, so that we can obtain the correct Einstein–Hilbert action after $d^n x$ integration, we need to have

$$\lambda = \left[\frac{\det g_{mn}}{\det g^{(0)}_{mn}} \right]^{-\frac{1}{d-2}} . \tag{13.49}$$

In the case of the original KK reduction, with $d = 4$ and $D = 5$, we get $g^{(5)}_{\mu\nu} = g^{(4)}_{\mu\nu} g_{55}^{-1/2}$. Defining

$$g_{55} \equiv \Phi^{2/3} , \tag{13.50}$$

we obtain $g^{(5)}_{\mu\nu} = \Phi^{-1/3} g^{(4)}_{\mu\nu}$. In order to also find the Maxwell action for $g_{\mu 5}$, we need to redefine it also as

$$g_{\mu 5} = B_\mu(\vec{x}, t) g_{55}. \tag{13.51}$$

Finally, that means that the full nonlinear ansatz for the original KK dimensional reduction is

$$g_{\Lambda\Sigma} = \Phi^{-1/3}(\vec{x},t)\begin{pmatrix} g_{\mu\nu}(\vec{x},t) & B_\mu(\vec{x},t)\Phi(\vec{x},t) \\ B_\mu(\vec{x},t)\Phi(\vec{x},t) & \Phi(\vec{x},t)\end{pmatrix}. \tag{13.52}$$

Of course, if we want to put $\Phi = 1$, writing this nonlinear ansatz is moot, but as we said, this is not a consistent truncation.

Important Concepts to Remember

- We can easily generalize electromagnetism and gravity to higher dimensions, just defining $G_N^{(D)}$ as the same coefficient in the Einstein–Hilbert action.
- Under KK dimensional reduction, we obtain $G_N^{(D)} = G_N^{(4)} V^{(n)}$, and we define $1/(8\pi G_N^{(D)}) = [M_{\rm Pl}^{(D)}]^{D-2}$.
- For $r \ll R$, we have for the gravitational acceleration $|\vec{g}(r)| \sim G_N^{(D)} M/r^{D-2}$, whereas for $r \gg R$, we have $|\vec{g}(r)| \sim G_N^{(D)}/[(2\pi R)^{D-4} r^2]$.
- For a KK dimensional reduction, the truncation can be inconsistent, due to terms in the action linear in the modes to be put to zero.
- The truncation becomes consistent if the integral of the Y_n harmonic with the other Y_0 harmonics gives zero, which happens in the torus case, as well as in the case when we keep *all* the singlets under the symmetry group G in the truncation.
- Sometimes nonlinearly redefining the fields, or equivalently, for a nonlinear KK reduction ansatz, can make the truncation consistent.
- To obtain the Einstein–Hilbert action from the same in the higher dimension, we need to nonlinearly redefine $g_{\mu\nu}^{(D)} = g_{\mu\nu}^{(4)} [\det g_{\mu\nu}/\det g_{\mu\nu}^{(0)}]^{-1/(d-2)}$.

Further reading: For consistent truncations of supergravity, see [14].

Exercises

(1) Calculate the gravitational potential of a thin shell of radius R and width $\Delta R \ll R$ and density ρ in 4 dimensions, in all relevant radial regions.
(2) Consider an insulating sphere in 4 dimensions (shell only), of surface charge density σ and radius R, situated between 2 parallel conducting plates, each at a distance $2R$ from the center of the sphere. Calculate the potential at the position of one of the conductors.
(3) Consider the Einstein–Hilbert action coupled nonminimally with the scalar ϕ, with action

$$S = \frac{M_{\rm Pl}^2}{2}\int d^4x\sqrt{-g}\left[R(1+\alpha\phi) - \beta\phi^{1+\frac{1}{n}}\right]. \tag{13.53}$$

Knowing the formula for the general Weyl rescaling in d dimensions

$$R[g_{\mu\nu}] = \Omega^{-2} R[g^E_{\mu\nu}] - 2(d-1) g_E^{\mu\nu} \nabla^E_\mu \nabla^\nu_E \ln\Omega - (d-2)(d-1) g_E^{\mu\nu} (\nabla^E_\mu \ln\Omega)(\nabla^E_\nu \ln\Omega) \,, \tag{13.54}$$

find the transformation that reduces the action to the Einstein–Hilbert action minimally coupled to a canonical dynamical scalar.

(4) Consider a charge q in between two perfectly conducting infinite conducting planes, situated both at a distance R from it. Calculate the electric potential on one of the conductors, at the minimum distance point from the charge. Write the result in arbitrary number of dimensions, for planes of codimension one.

(5) Prove that for the nonlinear ansatz for the original KK reduction from 5 dimensions, we obtain the Maxwell action for electromagnetism.

Chapter 14
$\mathcal{N} = 1$ Supergravity in 4 Dimensions

In this chapter we will study $\mathcal{N} = 1$ supergravity in 4 dimensions. Supergravity can be defined as a supersymmetric theory of gravity, or otherwise as a theory of local supersymmetry. In practice, we will use a bit of both definitions to find it. Before that however, we need to define what supersymmetry is. After that, we will count degrees of freedom, in order to see what supergravity models we can have. After defining the vielbein-spin connection formulation of gravity, within which supergravity is defined, we will finally define the model of $\mathcal{N} = 1$ supergravity.

14.1 Rigid Supersymmetry

The simplest model to understand supersymmetry in is the Wess–Zumino model in 2 dimensions. Indeed, supersymmetry is defined as a symmetry between bosons and fermions, but the symmetry must relate a degree of freedom with another. In the bosonic case, $\delta\phi^i = \epsilon\phi^i$ implicitly relates a bosonic degree of freedom to another, but it is kind of trivial, so we don't usually think about it that way. But in the case of supersymmetry, it will relate a bosonic degree of freedom with a fermionic one. As such, we can consider either on-shell supersymmetry, or off-shell supersymmetry, and the corresponding numbers of bosonic and fermionic degrees of freedom must match.

Here we will only describe on-shell supersymmetry. A Dirac fermion has $n = 2^{[d/2]}$ complex components. But in supersymmetry, one usually deals with Majorana fermions, which can be thought of as real. Thus an off-shell Majorana fermion has n real components, $n = 2$ for $d = 2$. On-shell, the Dirac equation $(\partial\!\!\!/ + m)\psi = 0$ is a matrix equation that relates half of the components with the other half, so an on-shell Majorana fermion has $n/2$ real components, 1 in the case of $d = 2$.

Therefore in $d = 2$ we have the simplest model, with a Majorana fermion ψ with a real on-shell degree of freedom, that can match the degree of freedom of a real

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_14

scalar ϕ. Then the simplest supersymmetric model is the free Wess–Zumino model, with the free action of a real scalar and a Majorana fermion,

$$S = -\frac{1}{2}\int d^2x[(\partial_\mu\phi)^2 + \bar{\psi}\partial\!\!\!/\psi]. \tag{14.1}$$

Since we have the mass dimensions $[d^2x] = -2$ and $[\partial_\mu] = 1$, the mass dimensions of the scalar and fermion are $[\phi] = 0$ and $[\psi] = 1/2$.

We must define the supersymmetry transformation rules. We take it as a definition that the scalar varies into the fermion times a parameter ϵ. For Lorentz invariance, ϵ must be also a spinor, and the variation must be

$$\delta\phi = \bar{\epsilon}\psi \equiv \bar{\epsilon}_\alpha\psi^\alpha \equiv \epsilon^\beta C_{\beta\alpha}\psi^\alpha. \tag{14.2}$$

A Majorana spinor satisfies the condition that the Majorana conjugate χ^C equals the Dirac conjugate $\bar{\chi}$, i.e.,

$$\chi^C \equiv \chi^T C = \bar{\chi} \equiv \chi^\dagger i\gamma_0. \tag{14.3}$$

Here C is the charge conjugation matrix, which can be defined generally in any dimension, but in 2 and 4 dimensions, is defined as an antisymmetric matrix that transforms γ_μ into $-\gamma_\mu^T$, i.e.,

$$C^T = -C; \quad C\gamma^\mu C^{-1} = -(\gamma^\mu)^T. \tag{14.4}$$

In dimensions other than 2 and 4, we can have either plus or minus for either of the two relations.

From the definition of supersymmetry, $\delta\phi = \bar{\epsilon}\psi$, we see that $[\epsilon] = -1/2$. If we want to write $\delta\phi$, it should be proportional to ϕ and ϵ. Now the Lorentz transformations match, but the dimensions don't match, we need to introduce an object of mass dimension 1. It could be ∂_μ, but we need to contract the μ index in order not to upset the Lorentz transformation property, so we multiply by $\partial\!\!\!/ = \gamma_\mu\partial_\mu$ instead, so that

$$\delta\psi = \partial\!\!\!/\phi\epsilon. \tag{14.5}$$

We now prove that the free WZ action is invariant under the supersymmetry (susy) transformation laws that we defined.

Majorana Spinor Identities

Before that however, we must prove some identities for Majorana spinor identities, valid both in 2 and 4 dimensions. There are more similar ones we could write, but those depend on dimension. We have

$$\bar{\epsilon}\chi = \bar{\chi}\epsilon \tag{14.6}$$

$$\bar{\epsilon}\gamma_\mu\chi = -\bar{\chi}\gamma_\mu\epsilon. \tag{14.7}$$

The first relation is proven as follows,

$$\bar{\epsilon}\chi = \epsilon^{\alpha} C_{\alpha\beta}\chi^{\beta} = -\chi^{\beta} C_{\alpha\beta}\epsilon^{\alpha} = +\chi^{\beta} C_{\beta\alpha}\epsilon^{\alpha} \equiv \bar{\chi}\epsilon \ , \tag{14.8}$$

where in the second equality the minus comes from permuting the order of the two fermions, and in the third we used $C^T = -C$. The second relation is proved similarly, by first multiplying the second equation defining the C matrix (14.4) by C, to find

$$C\gamma_{\mu} = -\gamma_{\mu}^T C = +\gamma_{\mu}^T C^T = +(C\gamma_{\mu})^T \ , \tag{14.9}$$

thus that $(C\gamma_{\mu})$ is symmetric. Then we have

$$\bar{\epsilon}\gamma_{\mu}\chi \equiv \epsilon^{\alpha}(C\gamma_{\mu})_{\alpha\beta}\chi^{\beta} = -\chi^{\beta}(C\gamma_{\mu})_{\alpha\beta}\epsilon^{\alpha} = -\chi^{\beta}(C\gamma_{\mu})_{\beta\alpha}\epsilon^{\alpha} \equiv -\bar{\chi}\gamma_{\mu}\epsilon. \tag{14.10}$$

We now return to the invariance of the action. Varying (14.1), we obtain

$$\begin{aligned} \delta S &= -\int d^2x \left[-\delta\phi\Box\phi + \frac{1}{2}\delta\bar{\psi}\partial\!\!\!/\psi + \frac{1}{2}\bar{\psi}\partial\!\!\!/\delta\psi \right] \\ &= -\int d^2x [-\delta\phi\Box\phi + \delta\bar{\psi}\partial\!\!\!/\psi] \ , \end{aligned} \tag{14.11}$$

where we have used the relation (14.6). Replacing the susy rules inside the variations, we find

$$\begin{aligned} \delta S &= -\int d^2x [-\bar{\epsilon}\psi\Box\phi + \bar{\psi}\partial\!\!\!/\partial\!\!\!/\phi\epsilon] \\ &= -\int d^2x [-(\bar{\epsilon}\psi)\Box\phi + (\bar{\psi}\epsilon)\Box\phi] \\ &= 0 \ , \end{aligned} \tag{14.12}$$

where in the second equality we used the relation

$$\partial\!\!\!/\partial\!\!\!/ = \gamma^{\mu}\gamma^{\nu}\partial_{\mu}\partial_{\nu} = \frac{1}{2}\{\gamma^{\mu}, \gamma^{\nu}\}\partial_{\mu}\partial_{\nu} = \partial^2 \equiv \Box. \tag{14.13}$$

In conclusion, we find that the action (14.1) is invariant under the susy rules *without using the equations of motion*, i.e., off-shell. That is however strange, since we have an on-shell supersymmetry only. But the point is that in general for a symmetry it is not enough to have an action that is invariant under it, we also need the algebra of the symmetries to close on the fields. In the case of usual, bosonic, symmetries, this is somewhat trivial, since the Lie algebra means that the commutator of two symmetries is another symmetry of the same type, so

$$[T_a, T_b] = f_{ab}{}^{c} T_c \Rightarrow [\epsilon_1^a T_a, \epsilon_2^b T_b] = (\epsilon_1^a f_{ab}{}^{c} \epsilon_2^b T_c) \Rightarrow [\delta_{\epsilon_1}, \delta_{\epsilon_2}] = \delta_{\epsilon_1 f \epsilon_2}. \tag{14.14}$$

This is usually satisfied. But in the case of supersymmetry, this is less obvious. The susy algebra that we need to satisfy is

$$\{Q^i_\alpha, Q^j_\beta\} = 2(C\gamma^\mu)_{\alpha\beta} P_\mu + \dots , \tag{14.15}$$

which means that we need to have on fields (multiplying by $\epsilon_1^\alpha \epsilon_2^\beta$, and considering that P_μ is represented by ∂_μ)

$$[\delta_{\epsilon_1}, \delta_{\epsilon_2}] = 2\bar{\epsilon}_2 \gamma^\mu \epsilon_1 \partial_\mu (+\dots) \tag{14.16}$$

It turns out that the algebra is realized on ϕ without the use of the equations of motion (off-shell), which is left as an exercise to prove. That is how we derive the algebra above. On the other hand, on the fermions, we need to use the equations of motion, i.e., it is on-shell, but we will not show it here.

14.2 Supergravity

We are now ready to define supergravity, as a supersymmetric theory of gravity, and a theory of local supersymmetry. In general, if we want to make a rigid symmetry local, we need to introduce a gauge field. For instance, if we have a complex scalar ϕ with action $S = -\frac{1}{2}\int d^d x |\partial_\mu \phi|^2$, invariant under $\phi \to e^{i\alpha}\phi$, we can make the invariance local, $\alpha \to \alpha(x)$, by introducing a gauge field, transforming as $\delta A_\mu = \partial_\mu \alpha$, and replace the derivative with the covariant derivative, $\partial_\mu \to D_\mu = \partial_\mu - i A_\mu$.

In the case of supersymmetry then, if we make it local by $\epsilon \to \epsilon(x)$, we need to introduce a "gauge field of local supersymmetry", "A^α_μ"$\to \psi^\alpha_\mu$, that has a spinor index α since the parameter ϵ does. This field ψ^α_μ is called the gravitino, and it is a spin 3/2 field (more precisely, on-shell we need to remove its gamma-trace, as we will see later). In fact, it turns out that the gravitino is also the superpartner of the graviton, so we can write the simplest $\mathcal{N} = 1$ supergravity just from the graviton and the gravitino.

Counting Degrees of Freedom

Before we construct supergravity, we want to describe the degrees of freedom in it. We have seen the counting of degrees of freedom of a spin 1/2 fermion, and the counting of the spin 1 gauge field is also familiar: off-shell, we have the d degrees of freedom, minus the one degree of freedom corresponding to gauge invariance, $\delta A_\mu = \partial_\mu \alpha$, giving $d - 1$ off-shell degrees of freedom. On shell, in the Lorenz gauge $\partial^\mu A_\mu = 0$, the equation of motion is simply the KG equation, $\Box A_\mu = 0$, which doesn't change the number of degrees of freedom, but the Lorenz condition itself is one condition, so on shell we have $d - 2$ degrees of freedom.

Graviton

1. *Off-shell*

For the graviton, the metric $g_{\mu\nu}$ is a symmetric matrix, so it has $d(d+1)/2$ components, but they are subject to the general coordinate transformation invariance, with parameter $\xi^{\mu}(x)$, i.e., we need to subtract d degrees of freedom, so we have $d(d-1)/2$ off-shell degrees of freedom.

2. *On-shell*

The linearized action for the graviton $h_{\mu\nu}$ ($g_{\mu\nu} = \eta_{\mu\nu} + 2\kappa_N h_{\mu\nu}$) coming from the Einstein–Hilbert action is the Fierz–Pauli action,

$$\mathcal{L} = \frac{1}{2}h^2_{\mu\nu,\rho} + h^2_{\mu} - h^{\mu}h_{,\mu} + \frac{1}{2}h^2_{,\mu} \,, \tag{14.17}$$

where

$$h_{\mu} \equiv \partial^{\nu}h_{\nu\mu}; \quad h \equiv h^{\mu}_{\mu} \,, \tag{14.18}$$

and, $\mu \equiv \partial_{\mu}$.

Just like in the case of the Lorenz gauge for electromagnetism, we can now choose a gauge, namely the *de Donder gauge*,

$$\partial^{\nu}\bar{h}_{\mu\nu} = 0 \,, \tag{14.19}$$

where

$$\bar{h}_{\mu\nu} \equiv h_{\mu\nu} - \frac{h}{2}\eta_{\mu\nu} \,, \tag{14.20}$$

in which the equation of motion becomes simply Klein-Gordon,

$$\Box\bar{h}_{\mu\nu} = 0. \tag{14.21}$$

We leave proving this as a simple exercise. The KG equation doesn't change the number of degrees of freedom, simply imposes $k^2 = 0$ in momentum space, but the de Donder gauge condition amounts to d local conditions, thus eliminating d degrees of freedom.

In conclusion, on-shell for the graviton we have $d(d-1)/2 - d = (d-1)(d-2)/2 - 1$ degrees of freedom, corresponding to a transverse ($d-2$ components only) symmetric traceless matrix.

Gravitino

1.*Off-shell*

Off-shell, $\psi_{\mu\alpha}$ has nd components, where $n = 2^{[d/2]}$. But it is also invariant under a gauge-like transformation, namely local supersymmetry (supergravity), which we will see that it is $\delta\psi_{\mu} = D_{\mu}\epsilon$, so it removes n components, leaving $n(d-1)$ components off-shell.

2.*On-shell*
Naively, we would say that, since $\psi_{\mu\alpha}$ is a gauge field, but also a fermion, on-shell we would have $n/2(d-2)$ components, where $n/2$ comes from the fact that the Dirac-like gravitino equation relates again half of the components to the other half.

However, there is a subtlety. As we already mentioned, the indices of $\psi_{\mu\alpha}$ mean that we make a product of the spin 1 and spin 1/2 representations, but that in fact decomposes as two irreducible representations,

$$1 \otimes 1/2 = 3/2 \oplus 1/2. \tag{14.22}$$

The point is that the gamma-trace, $\gamma^\mu \psi_\mu$, is a spin 1/2 fermion that transforms into itself, so lives in a separate irreducible representation, and only the gamma-traceless part represents the spin 3/2 representation. That means that the correct number of on-shell degrees of freedom of the gravitino is actually

$$\frac{n}{2}(d-2) - \frac{n}{2} = \frac{n}{2}(d-3). \tag{14.23}$$

We can now verify that in 4 dimensions, on-shell the graviton has $(d-1)(d-2)/2 - 1 = 3 \cdot 2/2 - 1 = 2$ degrees of freedom, and the gravitino has $2^{[d/2]}(d-3)/2 = 2^2 \cdot 1/2 = 2$ degrees of freedom also. That means that indeed, we can construct a $\mathcal{N} = 1$ supergravity model just from the graviton and the gravitino.

14.3 Vielbein-Spin Connection Formulation of General Relativity

But in order to define the supergravity model, we need to use the vielbein-spin connection formulation of general relativity, since there are fermions in the theory, that couple to gravity using the spin connection only.

The vielbein e^a_μ is a sort of square root of the metric $g_{\mu\nu}$, which is written as

$$g_{\mu\nu}(x) = e^a_\mu(x) e^b_\nu(x) \eta_{ab}. \tag{14.24}$$

Here μ, ν are "curved" indices, in the curved space, whereas a, b are "flat" indices, in a space tangent to the curved space at each point. As such, there is a *local Lorentz* symmetry acting on them, with Minkowski metric η_{ab}.

Therefore the local Lorentz (l.L.) and general coordinate (g.c.) transformations of the vielbein are

$$\begin{aligned} \delta_{l.L.} e^a_\mu(x) &= \lambda^a{}_b(x) e^b_\mu(x) \\ \delta_{g.c.} e^a_\mu(x) &= (\xi^\rho \partial_\rho) e^a_\mu + (\partial_\mu \xi^\rho) e^a_\rho. \end{aligned} \tag{14.25}$$

Here note that the local Lorentz transformation acts in the usual way on the index a, and the general coordinate transformation acts as on a vector on e_μ.

A note on the nomenclature. The term "vielbein" was originally defined in 4 dimensions only, where it was "vierbein" from the German "vier" = four and "bein" = leg. Then it was extended to ein-, zwei-, drei-bein (one, two, three legs in one, two, three dimensions), until finally it was generalized to "vielbein" for "viel" = many in an arbitrary dimension.

The *spin connection* is a kind of gauge field ("connection" in mathematical language) for curved space, specifically for the covariant derivative acting on spinors. The generator of Lorentz transformations in the spinor representation is $\frac{1}{4}\Gamma_{ab}$, so the covariant derivative on the spinors is defined as

$$D_\mu \psi = \partial_\mu \psi + \frac{1}{4}\omega_\mu^{ab}\Gamma_{ab}\psi. \tag{14.26}$$

But in order to obtain the usual general relativity, we cannot have both the vielbein e_μ^a and the spin connection ω_μ^{ab} be independent, we must write ω_μ^{ab} in terms of e_μ^a, otherwise we have too many degrees of freedom.

In fact, in terms of e_μ^a, we can check that we already have the correct number of degrees of freedom, since we have d^2 components, minus the d components of the general coordinate transformations defined by ξ^μ, minus the $d(d-1)/2$ components of the local Lorentz transformations, leading to the same $(d-2)(d-1)/2-1$ degrees of freedom.

The condition that we need to impose is known as the no-torsion constraint, or the "vielbein postulate". The torsion is defined as the antisymmetric part of the covariant derivative of the vielbein, $T_{\mu\nu}^a \equiv D_{[\mu}e_{\nu]}^a$, and in general the covariant derivative is

$$D_\mu e_\nu^a = \partial_\mu e_\nu^a - \Gamma_{\mu\nu}^\lambda e_\lambda^a + \omega_\mu^{ab} e_\nu^b \,, \tag{14.27}$$

but the Chistoffel symbol is symmetric in $(\mu\nu)$, so by taking the antisymmetric part we obtain

$$T_{\mu\nu}^a = D_{[\mu}e_{\nu]}^a = \partial_{[\mu}e_{\nu]}^a + \omega_{[\mu}^{ab}e_{\nu]}^b = 0. \tag{14.28}$$

The solution of this equation is an $\omega_\mu^{ab} = \omega_\mu^{ab}(e)$, which is left as an exercise to prove.

We can define the field strength of the "gauge field" ω_μ^{ab} (for the $SO(1,3)$ local Lorentz group) is

$$R_{\mu\nu}^{ab}(\omega) = \partial_\mu \omega_\nu^{ab} - \partial_\nu \omega_\mu^{ab} + \omega_\mu^{ac}\omega_\nu^{cb} - \omega_\nu^{ac}\omega_\mu^{cb}. \tag{14.29}$$

With this definition, this field strength, in the case the vielbein is $\omega = \omega(e)$, is just the Riemann tensor (in terms of the Christoffel tensor) with two indices flattened by the vielbein, so

$$R_{\mu\nu}^{ab}(\omega(e)) = R^\mu{}_{\nu\rho\sigma}(\Gamma(g(e)))e_\mu^a e^{-1\nu b}. \tag{14.30}$$

Since moreover, as a matrix we have

$$g = e \cdot \eta \cdot e \Rightarrow \det g = -(\det e)^2 \,, \tag{14.31}$$

in the Einstein–Hilbert action we can replace $\sqrt{-g}$ with $\det e$, and the Ricci scalar with the contraction of the field strength of ω, obtaining

$$S_{EH} = \frac{1}{16\pi G_N} \int d^4x (\det e) R^{ab}_{\mu\nu}(\omega(e))(e^{-1})^\mu_a (e^{-1})^\nu_b. \tag{14.32}$$

But since, as we can check,

$$\det e (e^{-1})^\mu_a (e^{-1})^\nu_b = \epsilon_{abcd} \epsilon^{\mu\nu\rho\sigma} e^c_\rho e^d_\sigma \,, \tag{14.33}$$

the Einstein–Hilbert action can be rewritten as

$$\begin{aligned} S_{EH} &= \frac{1}{2\kappa_N^2} \int d^4x \epsilon_{abcd} \epsilon^{\mu\nu\rho\sigma} R^{ab}_{\mu\nu}(\omega) e^c_\rho e^d_\sigma \\ &\equiv \frac{1}{2\kappa_N^2} \int \epsilon_{abcd} R^{ab}(\omega) \wedge e^c \wedge e^d \,, \end{aligned} \tag{14.34}$$

where on the last line we have written things in terms of forms, like $\omega^{ab} = \omega^{ab}_\mu dx^\mu$. Now we notice that in fact, we can consider $\omega = \omega(e)$, as it was implicit until now, or not, the result is the same. Indeed, considering now ω as an independent variable, we can vary with respect to it, and the resulting equation of motion is

$$\epsilon_{abcd} De^c \wedge e^d = 0 \Rightarrow De^c = 0 : \quad D_{[\mu} e^a_{\nu]} = T^a_{\mu\nu} = 0 \,, \tag{14.35}$$

which is the same as the vielbein postulate defining $\omega(e)$!

14.4 $\mathcal{N} = 1$ Supergravity in 4 Dimensions

To construct the action of $\mathcal{N} = 1$ supergravity, we need to add to the Einstein–Hilbert action for the graviton an action for the gravitino.

In flat space, Rarita and Schwinger wrote the action for a spin 3/2 field, in a general dimension

$$S_{RS} = -\frac{1}{2} \int d^d x \bar{\psi}_\mu \gamma^{\mu\nu\rho} \partial_\nu \psi_\rho. \tag{14.36}$$

In 4 dimensions, the action can be rewritten using the relation

$$i \epsilon^{\mu\nu\rho\sigma} \gamma_5 \gamma_\nu = \gamma^{\mu\rho\sigma} \,, \tag{14.37}$$

where

$$\gamma_5 = i\gamma_0\gamma_1\gamma_2\gamma_3 \ , \tag{14.38}$$

as

$$S_{RS,4d} = -\frac{i}{2}\int d^4x\epsilon^{\mu\nu\rho\sigma}\bar{\psi}_\mu\gamma_5\gamma_\nu\partial_\rho\psi_\sigma. \tag{14.39}$$

We can easily write the action in curved space, thus coupling it to the graviton, by putting $\det e$ in the volume and replacing $\partial_\mu \to D_\mu$, so

$$S_{RS} = -\frac{1}{2}\int d^dx(\det e)\bar{\psi}_\mu\gamma^{\mu\nu\rho}D_\nu\psi_\rho \ , \tag{14.40}$$

and in 4 dimensions also

$$S_{RS,4d} = -\frac{i}{2}\int d^4x\epsilon^{\mu\nu\rho\sigma}\bar{\psi}_\mu\gamma_5\gamma_\nu D_\rho\psi_\sigma. \tag{14.41}$$

The supergravity action is then simply the sum of the two terms,

$$S_{\mathcal{N}=1\ sugra} = S_{EH}(e,\omega) + S_{RS}(\psi_\mu,\omega). \tag{14.42}$$

The supersymmetry transformation laws are written remembering the fact that supergravity is a supersymmetric theory of gravity, and a theory of local supersymmetry. Being a supersymmetric theory of gravity, the vielbein, standing in for gravity, must vary into the gravitino, like we had $\delta\phi = \bar{\epsilon}\psi$ for the 2 dimensional WZ model. Now the index a on e^a_μ doesn't correspond to anything on the right hand side, so we must introduce the constant matrix γ^a to compensate for it, giving finally

$$\delta e^a_\mu = \frac{\kappa_N}{2}\bar{\epsilon}\gamma^a\psi_\mu. \tag{14.43}$$

Note that the factor of $\kappa_N/2$ is conventional, it could have been absorbed in the definition of ϵ. Since ψ_μ is considered as a gauge field of local supersymmetry, and the local Lorentz group is nonabelian, we expect a nonabelian gauge field transformation of the type $\delta\psi_\mu = D_\mu\epsilon$. Indeed, we can write

$$\delta\psi_\mu = \frac{1}{\kappa_N}D_\mu\epsilon \ , \tag{14.44}$$

where as before, we have

$$D_\mu\epsilon = \partial_\mu\epsilon + \frac{1}{4}\omega^{ab}_\mu\Gamma_{ab}\epsilon. \tag{14.45}$$

Now that we have defined the supergravity action (14.42) and the transformation rules, one can check whether the action is invariant under them, and find that it is indeed. However, there is a subtlety. If we consider the spin connection ω as a depen-

dent quantity, we cannot however consider $\omega = \omega(e)$ as in the absence of fermions. This would be inconsistent, i.e., in particular would not lead to a supersymmetric theory.

Indeed, now, if we consider ω as independent in (14.42), we see that the equation of motion is modified due to the ω in D_μ action on fermions, to $D_{[\mu} e^a_{\nu]} = \bar{\psi}\psi$ terms, so its solution is

$$\omega = \omega(e, \psi) = \omega(e) + \bar{\psi}\psi \text{ terms}. \tag{14.46}$$

This would be a *first order* formulation of supergravity. If we want to consider instead a *second order* formulation of supergravity, with a dependent ω, by consistency we need to consider $\omega = \omega(e, \psi)$ instead of $\omega = \omega(e)$.

Complete Transformation Laws of Supergravity

To finish, we summarize the complete set of transformation laws of $\mathcal{N} = 1$ supergravity in 4 dimensions.

The supersymmetry transformation laws are

$$\begin{aligned} \delta_S e^a_\mu &= \frac{\kappa_N}{2} \bar{\epsilon} \gamma^a \psi_\mu \\ \delta_S \psi_\mu &= \frac{1}{\kappa_N} D_\mu \epsilon. \end{aligned} \tag{14.47}$$

In a first order formulation, we would need also the transformation laws for ω^{ab}_μ, but we will not write them here.

The Einstein (general coordinate) transformation laws are

$$\begin{aligned} \delta_E e^a_\mu &= \xi^\nu \partial_\nu e^a_\mu + (\partial_\mu \xi^\nu) e^a_\nu \\ \delta_E \omega^{ab}_\mu &= \xi^\nu \partial_\nu \omega^{ab}_\mu + (\partial_\mu \xi^\nu) \omega^{ab}_\nu \\ \delta_E \psi_\mu &= (\xi^\nu \partial_\nu) \psi_\mu + (\partial_\mu \xi^\nu) \psi_\nu. \end{aligned} \tag{14.48}$$

Finally, the local Lorentz transformation laws are

$$\begin{aligned} \delta_{l.L.} e^a_\mu &= \lambda^a{}_b e^b_\mu \\ \delta_{l.L.} \omega^{ab}_\mu &= D_\mu \lambda^{ab} = \partial_\mu \lambda^{ab} + \omega^{ac}_\mu \lambda^{cb} - \omega^{bc}_\mu \lambda^{ca} \\ \delta_{l.L.} \psi_\mu &= -\frac{1}{4} \lambda^{ab} \gamma_{ab} \psi_\mu. \end{aligned} \tag{14.49}$$

Important Concepts to Remember

- Supergravity is a theory of local supersymmetry, or a supersymmetric theory of gravity.
- The gauge field of the local susy, and the superpartner of the graviton, is the gravitino ψ_μ.
- Supersymmetry is a symmetry between bosons and fermions, that matches the number of degrees of freedom.

- The susy transformation rules are of the type $\delta\phi = \bar{\epsilon}\psi$ and $\delta\psi = \partial\!\!\!/\phi\epsilon$.
- On-shell supersymmetry uses the equations of motion and matches degrees of freedom on-shell, and off-shell supersymmetry matches degrees of freedom off-shell.
- In the off-shell case, the algebra of transformations needs to be realized on the fields, and in the on-shell case this only happens up to equations of motion.
- The gravitino ψ_μ contains also a spin 1/2 on-shell representation, the gamma trace. The spin 3/2 part is the gamma-traceless part.
- One can formulate general relativity in terms of vielbein, a type of square root of the metric, with a flat index, in the tangent space at a local point, acted upon by local Lorentz transformations; and the spin connection, the "gauge field" appearing in the covariant derivative that acts on spinors.
- The Einstein–Hilbert action is written in terms of the field strength of ω and the vielbein e, and the equation of motion for ω gives the "vielbein constraint" that is solved by $\omega = \omega(e)$ and reduces it to the EH action.
- The $\mathcal{N}=1$ supergravity action is the sum of the EH action for gravity and the Rarita-Schwinger action for the gravitino.
- The supersymmetry rules are $\delta e^a_\mu = \kappa_N/2\bar{\epsilon}\gamma^a\psi_\mu$ and $\delta\psi_\mu = D_\mu\epsilon/\kappa_N$.
- In the second order formulation of supergravity, $\omega = \omega(e,\psi)$ is the solution of the equations of motion of the action with independent ω.

Further reading: See [15, 16] for more on supersymmetry and supergravity. For a recent complete work on supergravity see [17]. For a very good review of supergravity, see [18].

Exercises

(1) Check that we can represent the susy algebra on the scalar ϕ of the 2 dimensional WZ model without the use of the equations of motion.
(2) Check that the spin connection as a function of the vielbein,

$$\omega^{ab}_\mu(e) = \frac{1}{2}(\partial_\mu e^a_\nu - \partial_\nu e^a_\mu) - \frac{1}{2}e^{b\nu}(\partial_\mu e^a_\nu - \partial_\nu e^a_\mu) - \frac{1}{2}e^{a\rho}e^{b\sigma}(\partial_\rho e_{c\sigma} - \partial_\sigma e_{c\rho})e^c_\mu\,, \tag{14.50}$$

satisfies the no-torsion (vielbein) constraint $T^a_{\mu\nu} \equiv 2D_{[\mu}e^a_{\nu]} = 0$.
(3) Prove that by imposing the de Donder gauge, the Fierz–Pauli action gives the KG equation for $\bar{h}_{\mu\nu}$.
(4) Find the spectrum (field content) and check the number of on-shell degrees of freedom for $\mathcal{N}=2$ supergravity in 4 dimensions, as a sum of the $\mathcal{N}=1$ supergravity multiplet and another $\mathcal{N}=1$ multiplet that starts with the gravitino (gravitino multiplet). Repeat the same exercise in 3 dimensions. Can you guess off-shell fields that need to be added to have also matching of degrees of freedom off-shell (off-shell susy)?
(5) Show that at the linearized level, the Einstein–Hilbert action reduces to the Fierz–Pauli action.

Chapter 15
KK Compactification of Supergravity Models

In this chapter, we will study the KK compactification of supergravity models in higher dimensions.

Supergravity models in higher dimensions have:

- a graviton, described by the metric $g_{\mu\nu}$ (which can be written as $e^a_\mu \eta_{ab} e^b_\nu$), except for the coupling to fermions, which is written in terms of $\omega^{ab}_\mu(e)$.
- gravitino(s) $\psi^i_{\mu\alpha}$, one for each supersymmetry. Indeed, each supersymmetry ϵ^i takes us from the graviton to a different gravitino, $\delta e^a_\mu = \bar{\epsilon}^i \gamma^a \psi^i_\mu$.
- other fields: scalars ϕ^I, vectors A^I_μ, spinors λ^I_α, and also antisymmetric tensors $A_{\mu_1 \dots \mu_r}$.

The antisymmetric tensors are generalizations of the gauge fields (Maxwell fields) A_μ, that have also a field strength,

$$F_{\mu_1 \dots \mu_{r+1}} = (n+1) \partial_{[\mu_1} A_{\mu_2 \dots \mu_{r+1}]} \,, \tag{15.1}$$

and so satisfy a gauge invariance

$$\delta A_{\mu_1 \dots \mu_r} = \partial_{[\mu_1} \Lambda_{\mu_2 \dots \mu_r]}. \tag{15.2}$$

The gauge invariant action for the antisymmetric tensor field is

$$S = -\frac{1}{2(r+1)!} \int d^d x \sqrt{-\det g} F^2_{\mu_1 \dots \mu_{r+1}}. \tag{15.3}$$

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_15

15.1 KK Reduction

We have already seen how to do KK reduction of scalars, metrics and gauge fields, but let us start by reviewing the latter. The gauge field $A_M(\vec{x}, t; \vec{y})$ splits as $(A_\mu(\vec{x}, t; \vec{y}), A_m(\vec{x}, t; \vec{y}))$, where $M = (\mu m)$ and $\mu = 0, 1, 2, 3$, $m = 4, \ldots, 4 + n$. The A_μ components transform as a vector under $\Lambda_\mu{}^\nu \in SO(1, 3)$ and A_m as scalars, i.e.,

$$\begin{aligned} A'_\mu(\vec{x}', t'; \vec{y}) &= \Lambda_\mu{}^\nu A_\nu(\vec{x}, t; \vec{y}) \\ A'_m(\vec{x}, t; \vec{y}) &= A_m(\vec{x}, t; \vec{y}). \end{aligned} \tag{15.4}$$

The KK expansion (which is simply a parametrization of the general function on the compact space) is

$$\begin{aligned} A_\mu(\vec{x}, t; \vec{y}) &= \sum_{q, I_q} A_\mu^{q, I_q}(\vec{x}, t) Y_q^{I_q}(\vec{y}) \\ A_m(\vec{x}, t; \vec{y}) &= \sum_{q, I_q} A^{q, I_q}(\vec{x}, t) Y_m^{q, I_q}(\vec{y}). \end{aligned} \tag{15.5}$$

Antisymmetric Tensors

Now we can easily generalize to the antisymmetric tensors, where $A_{M_1 \ldots M_p}(\vec{x}, t; \vec{y})$ splits into a set of components

$$\{A_{\mu_1 \ldots \mu_r m_{r+1} \ldots m_p}(\vec{x}, t; \vec{y})\} , \tag{15.6}$$

where $p - r \neq n \equiv dim(K_n)$ and $r \leq 4$, since we cannot have more antisymmetrized indices than the dimension of the space.

The KK expansion in spherical harmonics is

$$A_{\mu_1 \ldots \mu_r m_{r+1} \ldots m_p}(\vec{x}, t; \vec{y}) = \sum_{q, I_q} a_{\mu_1 \ldots \mu_r}^{q, I_q}(\vec{x}, t) Y_{m_{r+1} \ldots m_p}^{q, I_q}(\vec{y}). \tag{15.7}$$

Fermions

For fermions, if we would dimensionally reduce, i.e., keep just the zero mode, on a torus, we would write

$$\lambda_A(\vec{x}, t; \vec{y}) = \lambda_\tau^i(\vec{x}, t) , \tag{15.8}$$

where the index A in the spinor representation of $SO(1, 3 + n)$ would split into (τ, i), with τ an index in the spinor representation of $SO(1, 3)$, and i just an index filling up A.

In general, on a $M_4 \times K_n$ space, for the dimensional reduction we would write

$$\lambda_A(\vec{x}, t; \vec{y}) = \lambda_\tau^I(\vec{x}, t) \eta_i^I(\vec{y}) , \tag{15.9}$$

where I is an index in the symmetry group G of the compact space K_n, and again A splits as $A = (\tau i)$. But moreover, now $\eta_i^I(\vec{y})$ on spaces with symmetries (symmetry group G) are *Killing spinors*. These are a kind of "square root" of a Killing vector.

A Killing vector (named after Wilhelm Killing), associated with an isometry of a space K_n, is a vector V_μ^{AB} that satisfies the Killing vector equation

$$D_{(\mu} V_{\nu)}^{AB} = 0. \tag{15.10}$$

A Killing spinor on a sphere satisfies the condition

$$D_\mu \eta_i^I = c(\gamma_\mu \eta^I)_i \ , \tag{15.11}$$

where c is usually taken to be $i/2$. It is a square root of a Killing vector in the sense that

$$V_\mu^{AB} = c' \bar{\eta}^I \gamma_\mu \eta^J (\gamma^{AB})_{IJ}. \tag{15.12}$$

Usually, c' is taken to be $-\mathrm{i}/8$.

On a more general space, a Killing spinor can be defined in supersymmetric theories as the condition that the variation of the gravitino is zero, leading to the covariant derivative of the spinor being proportional to some gamma matrices times some constant fields.

Note that we want to have spinor fields in both $4 + n$ dimensions and in 4 dimensions. But then the spin-statistics theorem says that these fields must both be anticommuting. However, then the reduction relation (15.9) means that the Killing spinors must be *commuting* spinors (that is fine, since they are thought of as constants, not fields).

The Killing vector V_μ^{AB} is associated with the isometries of the space, which as we said turn into gauge symmetries of the reduced theory. Specifically, *at the linearized level*, the KK reduction ansatz for the off-diagonal part of the metric is

$$g_{\mu m}(\vec{x}, t; \vec{y}) = B_\mu^{AB}(\vec{x}, t) V_m^{AB}(\vec{y}). \tag{15.13}$$

At the nonlinear level, we must multiply by a factor involving the compact space metric.

15.2 $\mathcal{N} = 1$ Supergravity Coupled to Matter

We consider models of $\mathcal{N} = 1$ supergravity in 4 dimensions, coupled to matter. We are interested only in $\mathcal{N} = 1$ models for phenomenology reasons. We eventually want our models to contain the Standard Model of particle physics, but then we cannot have $\mathcal{N} = 2$ supersymmetry, since in it, chiral fermions come in opposite pairs, whereas we know that low energy physics is chiral (contains chiral fermions).

So in 4 dimensions, we can only have $\mathcal{N} = 1$ unbroken supersymmetry if we are to obtain the Standard Model after supersymmetry breaking.

Wess–Zumino Multiplets

One type of matter multiplets coupled to supergravity are the Wess–Zumino or chiral multiplets, composed of a complex scalar field ϕ and a (Majorana) fermion λ. We have seen in the last chapter that in 2 dimensions, the WZ model contains a Majorana fermion and a real scalar. But a fermion in 4 dimensions has twice as many degrees of freedom as one in 2 dimensions (since $n = 2^{[d/2]}$ is the number of components of a Dirac fermion), whereas a scalar has the same ones. That means that in 4 dimensions we need two real scalars, combining into a complex scalar, in order to form a multiplet, but otherwise the multiplet still has spin content $(1/2, 0)$. Moreover, in the last chapter, we have shown how to write the *free* WZ multiplet, but we can write also an interacting one, that includes self-interactions.

The interactions of the chiral multiplet are encoded into a function called the *superpotential* $W(\Phi)$. It comes from the superspace formulation of supersymmetry, which will not be described here in detail, other than to say that a chiral superfield $\Phi(x, \theta)$ is expanded in the fermionic coordinate θ as

$$\Phi(x, \theta) = \phi(x) + \sqrt{2}\theta\lambda(x) + \theta\theta F(x) + \cdots , \tag{15.14}$$

where the dots represent terms dependent on the ones written, and $F(x)$ is an *auxiliary field* (with quadratic action with no derivatives). The superpotential W is a holomorphic function of the superfield Φ (so is independent on $\bar{\Phi}$). Then, the most general renormalizable model has the cubic superpotential

$$W = \lambda\Phi + \frac{m}{2}\Phi^2 + \frac{g}{3}\Phi^3. \tag{15.15}$$

However, if we consider a quantum effective theory, then W can be anything, leading to a non-renormalizable theory.

The potential for the scalars comes from solving the equation of motion for the auxiliary field F and replacing in the action F^2, giving, in rigid supersymmetry and in the case of canonical scalars (with canonical kinetic terms)

$$V = F^2 = \sum_i \left| \frac{\partial W}{\partial \phi^i} \right|^2 . \tag{15.16}$$

Because of its origin, this is generally called an "F term".

In this case, the scalar kinetic Lagrangian is the usual one,

$$\mathcal{L}_{\text{kin,sc.}} = -(D_\mu \phi^i)(D^\mu \phi^i)^\dagger , \tag{15.17}$$

and the fermionic terms in the Lagrangian are

$$\mathcal{L}_{\text{fermi}} = -i\bar{\psi}\gamma^{\mu}\partial_{\mu}\psi - \frac{1}{2}\frac{\partial^2 W}{\partial\phi^i\partial\phi^j}\psi_i\psi_j - \frac{1}{2}\frac{\partial^2 \bar{W}}{\partial\phi^{\dagger i}\partial\phi^{\dagger j}}\bar{\psi}_i\bar{\psi}_j. \tag{15.18}$$

But the chiral multiplets can also have nontrivial (yet still second derivative) kinetic terms. This is described by another function, known as the *Kähler potential* $K(\Phi, \bar{\Phi})$. Unlike the superpotential, this is not a holomorphic function. In its presence, the scalar and fermion kinetic terms are modified to

$$\mathcal{L}_{\text{kinetic}} = g_{i\bar{j}}[(D_{\mu}\phi^i)(D^{\mu}\phi^j)^* + \psi^i \not{D}\bar{\psi}^{\bar{j}}] \,, \tag{15.19}$$

where we have defined the metric in scalar field space (Kähler metric)

$$g_{i\bar{j}} = \frac{\partial^2 K}{\partial\Phi^i\partial\bar{\Phi}^{\bar{j}}}. \tag{15.20}$$

Gauge Multiplets

The next possibility to couple to supergravity is the gauge multiplet, composed of a gauge field A^a_{μ} (in the adjoint of some gauge group G) and a fermion λ^a_{α}, in the same representation, so having spin content $(1, 1/2)$. The multiplet is defined by a function $\mathcal{F}(\{\Phi\})$ depending on all the chiral scalar multiplets Φ. Defining in the case of dependence on adjoint scalars Φ^a the quantity

$$\mathcal{F}_{ab} = \frac{\partial^2\mathcal{F}}{\partial\Phi^a\partial\Phi^b} \,, \tag{15.21}$$

the action for the gauge multiplets becomes

$$-\frac{1}{4}\text{Re}[\mathcal{F}_{ab}(\phi)F^a_{\mu\nu}F^{b\mu\nu}] - \frac{1}{2}\text{Re}[\mathcal{F}_{ab}(\phi)\bar{\lambda}^a \not{D}\lambda^b] + \frac{\partial W}{\partial\phi^i}\left(\frac{\partial\mathcal{F}_{ab}}{\partial\phi^j}\right)^* \bar{\lambda}^a\lambda^b. \tag{15.22}$$

Here we have assumed that $\mathcal{F}_{ab}(\Phi)$ depends on all the scalars, ϕ^i, not just those in the adjoint representation.

Moreover, the presence of the gauge fields, coupled to all the scalars, means that we actually add another term to the scalar potential, coming from the auxiliary field in the vector multiplet. We have, in a certain gauge, $V^a = \cdots + \theta^2\bar{\theta}^2 D^a$, so the last component of the vector superfield V^a is another auxiliary field called D^a. This will again lead to a new term in the scalar potential, known as "D term",

$$\frac{1}{2}|D_a|^2 = \frac{1}{2}|\Phi^{\dagger i}(T_a)_{ij}\Phi^j|^2. \tag{15.23}$$

There could be in principle also the gravitino multiplet, with spin content $(3/2, 1)$, but those generally are associated with extra supersymmetry for the graviton, which as we said we want to avoid, so that is it for fields with spins ≤ 2.

Coupling Supergravity to Matter

To couple to supergravity, we must consider a finite $M_{\rm Pl}$ (decoupled gravity corresponds to $M_{\rm Pl} \to \infty$). We first define a kind of covariant derivative,

$$D_i = \frac{\partial}{\partial \phi^i} + \frac{1}{M_{\rm Pl}^2} \frac{\partial K}{\partial \phi^i} \,, \tag{15.24}$$

in which case the potential for the scalars coupled to $\mathcal{N} = 1$ supergravity is

$$V = |F|^2 + |D^a|^2 = e^{\frac{K}{M_{\rm Pl}^2}} \left[\sum_{i\bar{j}} (g^{-1})^{i\bar{j}} D_i W (D_j W)^* - \frac{3}{M_{\rm Pl}^2} |W|^2 \right] + \frac{1}{2} (\mathcal{F}^{-1})^{ab} \left(\frac{\partial K}{\partial \phi^i} (T_a)_{ij} \phi^j \right) \left(\frac{\partial K}{\partial \phi^k} (T_b)_{kl} \phi^l \right) . \tag{15.25}$$

The gauge action stays unchanged,

$$\mathcal{L}_{\rm gauge} = -\frac{1}{4} {\rm Re}[\mathcal{F}_{ab}(\phi) F^a_{\mu\nu} F^{b\mu\nu}] \,, \tag{15.26}$$

whereas the gaugino (superpartner of the gauge field) action gets modified as well, to

$$\mathcal{L}_{\rm gaugino} = -\frac{1}{2} {\rm Re}[F_{ab}(\phi) \bar{\lambda}^a \not{D} \lambda^b] + \frac{e^{\frac{K}{2M_{\rm Pl}^2}}}{2} {\rm Re} \left[\sum_{i\bar{j}} (g^{-1})^{i\bar{j}} D_W \left(\frac{\partial \mathcal{F}_{ab}}{\partial \phi^j} \right)^* \bar{\lambda}^a \lambda^b \right] . \tag{15.27}$$

We have now described the most general action for $\mathcal{N} = 1$ supergravity coupled to matter.

15.3 Supergravity Compactifications

Supergravity Compactification on Tori

If we compactify supergravity theories on tori, we obtain abelian theories in 4 dimensions. Indeed, the torus $K_n = T^n$ has only abelian symmetries (of course, it has the nonabelian Lorentz symmetry $SO(n)$, but we are referring to isometries here, which are associated to gauge fields under dimensional reduction; $SO(n)$ gives only the usual global symmetry by dimensional reduction). Indeed, we saw that there wasn't a Killing spinor for the torus, or rather, we could think of the trivial (abelian) Killing spinors $\eta_i^I = \delta_i^I$.

Then in principle we have fields of the same spin coming from various different fields in the higher dimension. For instance, a vector field can come from $g_{\mu m}$, as well

as from components of a gauge field A_M in the higher dimension, or of antisymmetric tensors, components $A_{\mu m_1 \ldots m_r}$. But all these different fields will combine to form a multiplet of some *global* symmetry group in the KK reduced theory. As far as I know, there is no systematics to find the global symmetry group G, it is trial and error (case by case): we see what fields we get, and we try to fit them into representations of some symmetry group. For instance, the maximal $\mathcal{N} = 8$ supergravity in 4 dimensions has a global $SO(8)$ symmetry that transforms the gauge fields, but also a local composite symmetry $SU(8)$ and a global symmetry E_7 transforming the scalars. So the abelian gauge fields in general transform in some global symmetry group.

Nontrivial Compactifications and Gauged Supergravity

If we dimensionally reduce on a nontrivial space instead, like for instance a sphere $K_n = S^n$, the abelian Killing spinors of the torus turn into the nonabelian Killing spinors of K_n, and correspondingly the abelian gauge fields transforming in a *global* symmetry group G now turn into nonabelian (non-commuting) gauge fields of parts of, or all of the group G. This corresponds to making part of, or all the symmetry G local, or *gauging* it, with a nontrivial coupling constant g. The result is a *gauged supergravity*, and the ansatz needed to obtain it via dimensional reduction will be *nonlinear reduction ansatz*.

In the gauged supergravity, the susy transformation law for the gravitinos is modified by the addition of coupling constant (g) terms, to

$$\delta\psi^i_\mu = D_\mu(\omega(e,\psi))\epsilon^i + g\gamma_\mu\epsilon^i + gA_\mu\epsilon^i. \tag{15.28}$$

We see that there is a constant term in the transformation law, which is a nonlinear term. But a usual transformation law is linear, a field varying into another (times ϵ), and a nonlinearity like a constant term usually signifies the breaking of the symmetry, perhaps spontaneously. Indeed, in the Higgs effect for instance, $\delta\phi^i = M^i{}_j\phi^j\epsilon$ acquires a constant term under the split into VEV plus fluctuation, $\delta\varphi^i = M^i{}_j(v+\varphi)^j\epsilon$, that signals the breaking of the symmetry. However what happens now is that the action is still supersymmetric, but we need to add a *constant* term in the gravity part, that depends on g, i.e., a *cosmological constant* term $\int \det e\,\Lambda$. But it turns out that the sign of Λ is fixed to be negative, $\Lambda < 0$, in order to have supersymmetry.

Therefore gauged supergravity is AdS supergravity. And we can only get a de Sitter background by breaking supersymmetry (and a Minkowski background either by breaking susy, or by having an ungauged model).

15.4 11 Dimensional Supergravity

It turns out that all supergravities can be obtained from a single, unique, $\mathcal{N} = 1$ supergravity in 11 dimensions, by KK dimensional reduction and the use of symmetries. There doesn't seem to be a theorem about this, just an observation verified in

all examples. The unique supergravity in 11 dimensions has fields $g_{\mu\nu}$, $\psi_{\mu\alpha}$ (graviton and gravitino), but now, unlike in 4 dimensions, we need to add the antisymmetric 3-form field $A_{\mu\nu\rho}$. We can verify (by generalizing the counting of degrees of freedom to include antisymmetric tensors) that we have the same number of bosonic and fermionic degrees of freedom. The action is

$$\mathcal{L} = -\frac{e}{2}R(e,\omega) - \frac{e}{2}\bar{\psi}_\mu \Gamma^{\mu\nu\rho} D_\nu \psi_\rho - \frac{e}{48}F^2_{\mu\nu\rho} - \frac{1}{6}\epsilon^{\mu_1\ldots\mu_{11}} A_{\mu_1\mu_2\mu_3} F_{\mu_4\ldots\mu_7} F_{\mu_8\mu_{11}} \ , \tag{15.29}$$

and the field strength of the antisymmetric tensor is

$$F_{\mu\nu\rho\sigma} = 24\partial_{[\mu} A_{\nu\rho\sigma]}. \tag{15.30}$$

The first three terms are kinetic (Einstein–Hilbert, Rarita–Schwinger and Maxwell) and the last is a Chern–Simons term for $A_{\mu\nu\rho}$.

The reason that this supergravity is unique is the following. In 4 dimensions, the maximum amount of supersymmetry that doesn't involve spins greater than 2 is 8, since a supersymmetry changes the helicity by 1/2, and so by starting with helicity -2 (the minimum possible with spins up to 2) and acting 8 times, we reach $+2$. Otherwise, we would get helicities (and hence spins) higher than 2. We want to have spins ≤ 2, since there is no known interacting theory of a finite number of spins higher than 2 in flat space. Of course, string theory corresponds to an infinite tower of interacting higher spins, and there is the so-called Vasiliev theory of higher spins living in AdS background, but for a finite number of fields in flat space, there is no known interacting theory.

But then, the $\mathcal{N} = 8$ theory, with 8 gravitinos, can lift up only until 11 dimensions, since there the 8 minimal spinors of 4 dimensions (the gravitinos) combine to form a minimal spinor of 11 dimensions. In higher dimensions, we would not get a complete spinor, so we would break Lorentz invariance. That gives the 11 dimensional theory its uniqueness. On the other hand, not all supergravities are obtained by compactification alone. The simplest example is compactification on a circle, which gives the massless type IIA supergravity. But there is also the type IIB supergravity, obtained by a symmetry called T-duality.

The 11 dimensional supergravity has as its backgrounds preserving the maximal amount of supersymmetry 11 dimensional Minkowski space Min_{11}, the spaces $AdS_4 \times S^7$, $AdS_7 \times S^4$, as well as the pp waves obtained as limits ("Penrose limits") of these. The fact that there is no other maximally supersymmetric background is actually a theorem. Note that the $AdS_4 \times S^7$ solution, called the Freund–Rubin ansatz (with $F_{\mu\nu\rho\sigma} \propto \epsilon_{\mu\nu\rho\sigma}$, for μ, ν, ρ, σ in 4 dimensions) or spontaneous compactification, raised the hope that the theory itself will choose a compactified space with only 4 dimensions non-compact. In this solution however, the cosmological constant is related to the radius of the compact space, so that is clearly ruled out experimentally, and it was then proven that we cannot deform this situation, and always the cosmological constant will be related to the parameters of the compact space, in this "spontaneous compactification" type ansatz.

Important Concepts to Remember

- Supergravity models have gravitons, gravitinos (as many as supersymmetries), scalars, vectors, spinors and antisymmetric tensors.
- Fermions KK reduce to lower dimensional fermions times commuting Killing spinors.
- Killing spinors are square roots of Killing vectors, associated with the isometries of the space.
- The most general coupling of supergravity with matter in the $\mathcal{N} = 1$ context is defined by the superpotential and Kähler potential of chiral multiplets, and the $\mathcal{F}$ coupling function for vector multiplets.
- KK reduction of supergravity models on tori gives ungauged supergravity models, with only nonabelian global symmetries.
- KK reduction of supergravity models on nontrivial spaces (with nonabelian symmetries) gives gauged supergravity, where part or all of the global symmetry is gauged.
- Gauged supergravity is AdS supergravity, and we can only get de Sitter by breaking susy, and Minkowski by breaking susy or by ungauged models.
- All models can be obtained from the unique 11 dimensional supergravity, with fields $g_{\mu\nu}$, $\psi_{\mu\alpha}$ and $A_{\mu\nu\rho}$.

Further reading: See [13] for the KK reduction of supergravity, as well as [18]. For supergravity coupled to matter in a modern formulation, see [17, 19].

Exercises

(1) Consider a chiral multiplet, with the general renormalizable superpotential

$$W = \lambda\Phi + \frac{m}{2}\Phi^2 + \frac{g}{3}\Phi^3 \,, \tag{15.31}$$

coupled to supergravity, with Kähler potential

$$K = -3\ln(\Phi + \bar{\Phi}). \tag{15.32}$$

Calculate the scalar potential V.

(2) Consider the KK reduction of 11 dimensional supergravity on a circle S^1, with metric ansatz

$$ds^2 = G^{(1)}_{MN} dx^M dx^N = e^{-\frac{2\Phi}{3}} G^{S,(10)}_{\mu\nu} dx^\mu dx^\nu + e^{\frac{4\Phi}{3}} (dx^{10} + A_\mu dx^\mu)^2. \tag{15.33}$$

Calculate the field content in 10 dimensions, and the kinetic terms in the case Φ is constant.

(3) If

$$Y^A = \frac{1}{4}\bar{\eta}^I \gamma_5 \eta^J (\gamma^A)_{IJ} \tag{15.34}$$

for a 4-sphere S^4, show that

$$V_\mu^{AB} = Y^{[A} D_\mu Y^{B]} \tag{15.35}$$

is the Killing vector.
(4) Calculate the field content of the dimensional reduction of $\mathcal{N} = 1$ supergravity on a T^4 torus, down to 7 dimensions. How many supersymmetries are there in 7 dimensions?

Chapter 16
The Relativistic Point Particle

In this chapter, we will study the relativistic point particle, from the point of view of its "worldline", i.e., its path as it moves in spacetime. The purpose it to put it in a way that we can then generalize, to define string theory. Indeed, if we only think clearly enough about some mechanism for the particle, we can then easily do apply it to string theory.

The "worldline formalism" is a *first quantized* way to describe the relativistic quantum mechanics. The formalism of quantum field theory, sometimes called "second quantization", since we can think of it as a quantization of the wave function, is the standard way to describe relativistic quantum mechanics, but anything we can do with it, we can in principle do in the first quantized worldline formalism.

We start with writing the action for the particle, on its worldline.

The action for a nonrelativistic, free, point particle is

$$S = \int L dt = \int dt \, \frac{m\vec{v}^2}{2} = \int dt \, \frac{m\dot{\vec{x}}^2}{2} \, , \tag{16.1}$$

since $L = T - U$, and $U = 0$. We see that in the final form, it is written as an action on the worldline, defined by $x = x(t)$.

16.1 Relativistic Second Order Action

For a relativistic point particle, we need to find an action that reduces to the above in the nonrelativistic limit. As we know, the action is obtained by replacing dt with the relativistic proper time $d\tau$, with

$$d\tau^2 = -\eta_{\mu\nu} dx^\mu dx^\nu \, , \tag{16.2}$$

and the kinetic energy with minus the rest energy mc^2, i.e.,

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_16

$$S = -m(c^2) \int d\tau. \tag{16.3}$$

But we need to rewrite it in terms of the trajectory of the free particle, $X^\mu = X^\mu(\tau)$. From the definition of τ, by dividing both sides with $d\tau^2$, we see that

$$\frac{dX^\mu}{d\tau}\frac{dX^\nu}{d\tau}\eta_{\mu\nu} = -1. \tag{16.4}$$

This allows us to rewrite the action of the free particle as (writing $\dot{X}^\mu = dX^\mu/d\tau$)

$$S_1 = -m(c^2) \int d\tau \sqrt{-\dot{X}^\mu \dot{X}^\nu \eta_{\mu\nu}}. \tag{16.5}$$

In the nonrelativistic limit, *on the path* $X^i = X^i(t)$ *of the free particle*, we have

$$d\tau^2 = dt^2 - \frac{d\vec{x}^2}{c^2} = dt^2\left(1 - \frac{v^2}{c^2}\right), \tag{16.6}$$

so the action becomes

$$S_1 \simeq -mc^2 \int dt \sqrt{1 - \frac{v^2}{c^2}} \simeq \int dt \left[-mc^2 + \frac{m\vec{v}^2}{2}\right], \tag{16.7}$$

which is the nonrelativistic action plus a constant term.

Symmetries

The action S_1 is invariant under *reparametrizations*, that is, we can change the parametrization of the worldline, from the proper time τ to any other function $\tau' = \tau'(\tau)$. Under it, the coordinates X^μ are invariant (transform as scalars on the worldline), i.e., $X'^\mu(\tau'(\tau)) = X^\mu(\tau)$, since the reparametrization of the line doesn't modify the path itself. We write

$$\frac{dX^\mu}{d\tau} = \frac{dX^\mu}{d\tau'}\frac{d\tau'}{d\tau} = \frac{dX'^\mu}{d\tau'}\frac{d\tau'}{d\tau}, \tag{16.8}$$

so by plugging it into S_1, we obtain

$$\begin{aligned} S_1 &= -m \int d\tau \sqrt{-\eta_{\mu\nu}\frac{dX^\mu}{d\tau}\frac{dX^\nu}{d\tau}} \\ &= -m \int d\tau \frac{d\tau'}{d\tau}\sqrt{-\eta_{\mu\nu}\frac{dX'^\mu}{d\tau'}\frac{dX'^\nu}{d\tau'}} \\ &= -m \int d\tau' \sqrt{-\eta_{\mu\nu}\frac{dX'^\mu}{d\tau'}\frac{dX'^\nu}{d\tau'}}. \end{aligned} \tag{16.9}$$

Equations of Motion

The variation of the action S_1 gives

$$\begin{aligned}\delta S_1 &= -m\int d\tau\delta\left(\sqrt{-\dot{X}^\mu\dot{X}^\nu\eta_{\mu\nu}}\right)\\ &= -m\int d\tau\left[-\frac{\eta_{\mu\nu}\dot{X}^\mu\delta\dot{X}^\nu}{\sqrt{-\dot{X}^\rho\dot{X}_\rho}}\right]\\ &= -m\int d\tau\frac{d}{d\tau}\left[\frac{\eta_{\mu\nu}\dot{X}^\nu}{\sqrt{-\dot{X}^\mu\dot{X}^\nu\eta_{\mu\nu}}}\right]\delta X^\mu + \delta X^\mu m\frac{dX_\mu}{d\tau}\Big|_{\tau_i}^{\tau_f}\ ,\end{aligned} \tag{16.10}$$

where in the third equality we have used partial integration. Considering that $-\dot{X}^2 = 1$ and defining the momentum of the particle

$$p^\mu = mu^\mu = m\frac{dX^\mu}{d\tau}\ , \tag{16.11}$$

the equation of motion, coming from putting to zero the bulk of the variation δS_1, is

$$\frac{dp_\mu}{d\tau} = 0\ , \tag{16.12}$$

whereas the boundary condition, coming from the vanishing of the boundary term, is

$$\delta X^\mu p_\mu(\tau_f) - \delta X^\mu p_\mu(\tau_i) = 0. \tag{16.13}$$

The equation of motion is the equation of the free particle, and in this case it implies also the vanishing of the boundary term. The solution looks somewhat trivial, in the sense that it is the straight line of a free particle. But if we consider the motion in a *curved* spacetime, with $\eta_{\mu\nu}$ replaced with a general metric $g_{\mu\nu}$, we obtain instead the motion on a geodesic in the curved spacetime.

This is now nontrivial, and it corresponds to the interaction of a particle with the gravitational field. It is part of a more general situation, where background fields (like the metric here) appearing in the particle action correspond to interaction effects of the particle with the fields.

As another example, in order to couple to a background gauge field (gauge field), we add to the action the coupling to a current, $\int A_\mu j^\mu$. But normally, in the quantum field theory formalism (second quantization), this is a function of the spacetime, $\int d^4x A_\mu(X) j^\mu(X)$, and the current is a spatial delta function centered on the path of the particle,

$$j^\mu(X^\rho) = q\frac{dX^\mu}{d\tau}\delta^3(X^\rho - X^\rho(\tau)). \tag{16.14}$$

That means that the coupling becomes

$$\int d^4x A_\mu(X^\rho) j^\mu(X^\rho) = \int d^4x A_\mu(X^\rho) q \frac{dX^\mu}{d\tau} \delta^3(X^\rho - X^\rho(\tau)) = \int d\tau A_\mu(X^\rho(\tau)) \left(q \frac{dX^\mu(\tau)}{d\tau} \right). \tag{16.15}$$

16.2 Quantum Field Theory in the Worldline Formalism

Having defined a particle action and its coupling to the background fields, we can define the formalism of first quantization, i.e., the worldline formalism. This is somewhat unusual and is seldom taught in quantum field theory, since it is a bit cumbersome, and it works mostly in simple theories, like scalar theories or at most QED. Instead of starting from a field theory action, and deriving Feynman rules, and from them the Feynman diagrams, we must define the Feynman rules by hand.

- The propagator (from x to y) is defined from the particle action, as we will see shortly.
- The vertex factor at x and y, instead of being derived from the field theory action, must now be postulated. We have many ways of defining the vertex factor, corresponding to many possible (consistent) quantum field theories. In string theory however, as we will see, the difference is that the vertex is not defined ad hoc, but is completely constrained by the quantum consistency of the theory (in particular, by conformal invariance) and as a such, there is a unique result, corresponding to the fact that there is a unique consistent string theory.
- Finally, we must define rules about how to integrate the points. Here, it is rather obvious, namely $\int d^4x \int d^4y$, but in string theory it is less so. In fact, there are many subtleties for integration, related to possible overcounting and singularities, that make the general case very difficult.

Propagator from Action

To find the *spacetime* propagator between x and y, we need the *worldline* action.

We first note that we can write, for an operator as well as for a constant (number),

$$\langle y|\Box^{-1}|x\rangle = \int_0^\infty d\tau \langle y|e^{-\tau\Box}|x\rangle. \tag{16.16}$$

We then note that the Hamiltonian for a nonrelativistic particle of mass $m = 1/2$ is

$$H = \frac{\vec{p}^2}{2m} = \frac{\Box}{2m} = \Box. \tag{16.17}$$

That means that the transition amplitude for (Euclidean) time separation τ obtained with this Hamiltonian can be written as a path integral,

$$\langle y|e^{-\tau H}|x\rangle = \int_x^y \mathcal{D}X(t)e^{-\frac{1}{4}\int_0^\tau dt\dot{X}^2} , \tag{16.18}$$

so that the particle propagator,

$$\left\langle y|\frac{1}{\Box}|x\right\rangle = \int_0^\infty d\tau\langle y|e^{\tau\Box}|x\rangle = \int_0^\infty d\tau\int_x^y \mathcal{D}X(t)e^{-\frac{1}{2}S_P} , \tag{16.19}$$

where the (massless) particle action is, as we will see shortly,

$$S_P = -\frac{1}{2}\int_0^\tau dt\ \dot{X}^2. \tag{16.20}$$

In conclusion, once we know the action, we can calculate the propagator (and then we need to define the vertices and rules for integration).

16.3 First Order Particle Action

In order to have the propagator (and in general, in order to quantize), we need an action. But the action S_1 is highly nonlinear in $X^\mu(\tau)$, so it is hard to quantize, and it gives a complicated propagator.

Instead of it, we must define a first order formulation of the action, which is quadratic in the fields (and perhaps in derivatives as well). To general procedure on how to achieve that is to introduce some auxiliary fields. Here, the auxiliary field must be an independent (intrinsic) "worldline metric" in one dimension, $\gamma_{\tau\tau}(\tau)$. Better yet, consider the "vielbein" (the name is however inappropriate now, since we are in one dimension, so we should rather use "einbein")

$$e(\tau) = \sqrt{-\gamma_{\tau\tau}(\tau)} > 0. \tag{16.21}$$

Then we note that in one dimension we have $\sqrt{-\det\gamma} = e(\tau)$, and

$$\sqrt{-\det\gamma}\gamma^{\tau\tau} = -e^{-1}(\tau). \tag{16.22}$$

We could continue by trying to find the action that gives the second order form S_1, but it is easier instead to write the action as a general quadratic action for the scalar fields $X^\mu(\tau)$. One invariant term is the scalar kinetic term $-1/2\int d\tau\sqrt{-\det\gamma}\gamma^{\tau\tau}\frac{dX^\mu}{d\tau}\frac{dX^\nu}{d\tau}\eta_{\mu\nu}$, and another is the cosmological constant term $-\int d\tau\sqrt{-\det\gamma}m^2$. The action then becomes

$$S_P = \frac{1}{2}\int d\tau \left[e^{-1}(\tau)\frac{dX^\mu}{d\tau}\frac{dX^\nu}{d\tau}\eta_{\mu\nu} - em^2 \right] , \tag{16.23}$$

and is invariant under reparametrizations. Indeed, reparametrization invariance means

$$e'(\tau')d\tau' = e(\tau)d\tau \Rightarrow \frac{e'^{-1}(\tau')}{d\tau'} = \frac{e^{-1}(\tau)}{d\tau}. \tag{16.24}$$

The action we have obtained indeed is equivalent to S_1, since the equation of motion for $e(\tau)$ is

$$-\frac{1}{e^2}\dot{X}^2 - m^2 = 0 , \tag{16.25}$$

with the solution

$$e = \frac{1}{m}\sqrt{-\eta_{\mu\nu}\dot{X}^\mu\dot{X}^\nu}. \tag{16.26}$$

Replacing this solution in S_P, we obtain

$$S_P = -m\int d\tau\sqrt{-\eta_{\mu\nu}\dot{X}^\mu\dot{X}^\nu} = S_1. \tag{16.27}$$

Therefore S_P is the first order form of S_1; they are classically equivalent, though quantum mechanically they can in principle be different.

But S_P is much simpler, and in particular is quadratic in fields, so we can base our quantization on it. Moreover, we are really interested in the massless limit $m \to 0$, and the nonlinear action S_1 doesn't have a good $m \to 0$ limit (it vanishes under it), whereas S_P does: it is

$$S_{P,m\to 0} = \frac{1}{2}\int d\tau \left[e^{-1}(\tau)\dot{X}^\mu\dot{X}^\nu\eta_{\mu\nu} \right]. \tag{16.28}$$

The action still has reparametrization invariance ("gauge invariance"), under the transformation $e'(\tau')d\tau' = e(\tau)d\tau$. That means that we can fix a gauge by performing such a transformation, and in this gauge $e(\tau)$ is anything that we want. The simplest choice is $e(\tau) = 1$, for which we have the simplest quadratic action

$$S_P = \frac{1}{2}\int d\tau\ \dot{X}^\mu\dot{X}^\nu\eta_{\mu\nu}. \tag{16.29}$$

Equations of Motion and Constraints

The equations of motion of S_P in the gauge $e(\tau) = 1$ is

$$\frac{d}{d\tau}\left(\frac{dX^\mu}{d\tau}\right) = 0. \tag{16.30}$$

But since now we have fixed a gauge, we need to impose the equation of motion of the field that was fixed, $e(\tau)$, as a constraint. This is similar to what we do in electromagnetism, in the gauge $A_0 = 0$, where we need to impose the Gauss constraint, i.e., the would-be equation of motion of A_0.

Now the constraint is

$$-e^2 \frac{\delta S}{\delta e} = \frac{\delta S}{\delta e^{-1}} = 0 \Rightarrow \frac{dX^\mu}{d\tau}\frac{dX^\nu}{d\tau}\eta_{\mu\nu} = 0. \tag{16.31}$$

But this gives the energy-momentum tensor

$$T_{\gamma\gamma} \equiv \frac{2}{\sqrt{-\det\gamma}}\frac{\delta S}{\delta \gamma^{\tau\tau}} = -\frac{2}{e}\frac{\delta S}{\delta e^{-2}} = e^2 \frac{\delta S}{\delta e}\ , \tag{16.32}$$

so the constraint is the vanishing of the energy-momentum tensor,

$$T_{\gamma\gamma} \equiv T = 0. \tag{16.33}$$

16.4 Quantization of the Particle in the Light-Cone Gauge

We could continue with the quantization of the particle in the gauge $e = 1$, but we will leave that as an exercise. Instead, we will quantize it in the light-cone gauge.

Considering light-cone coordinates

$$X^\pm = \frac{t \pm X^1}{\sqrt{2}}\ , \tag{16.34}$$

we fix a gauge for reparametrization invariance by imposing a light-cone gauge condition

$$X^+ = \frac{1}{m^2}p^+\tau. \tag{16.35}$$

This fixes the reparametrization for τ, but it is not the usual proper time, nor is it the same as in the $e = 1$ gauge. Indeed, now we obtain $\dot{X}^+ = p^+/m^2$, but on the other hand we have in general (from S_1, or from S_P by eliminating e)

$$p_\mu = \frac{\partial L}{\partial \dot{X}^\mu} = \frac{m\dot{X}_\mu}{\sqrt{-\dot{X}^2}}\ , \tag{16.36}$$

so that $p^2 + m^2 = 0$. By comparing the two forms for p^μ, we obtain

$$p^+ = \frac{m\dot{X}^+}{\sqrt{-\dot{X}^2}} = m^2\dot{X}^+ \Rightarrow \dot{X}^2 = -\frac{1}{m^2}. \tag{16.37}$$

Replacing back in the definition of p_μ, we find

$$p^\mu = m^2 \dot{X}^\mu. \tag{16.38}$$

But as we saw, $e = \sqrt{-\dot{X}^2}/m$, so that now we find

$$e = \frac{1}{m}\sqrt{-\dot{X}^2} = \frac{1}{m^2}. \tag{16.39}$$

On the other hand, the mass-shell condition $p^2 + m^2 = 0$ becomes in light-cone coordinates

$$-2p^+p^- + p^I p^I + m^2 = 0\ , \tag{16.40}$$

where p^I is the transverse momentum, solved by

$$p^- = \frac{1}{2p^+}(p^I p^I + m^2). \tag{16.41}$$

Then p^-, as well as its canonically conjugate x^+, is a dependent quantity. The independent variables, from $p^\mu = m^2 \dot{X}^\mu$, are found to be

$$\begin{aligned} \frac{dX^-}{d\tau} &= \frac{p^-}{m} \Rightarrow X^-(\tau) = x_0^- + \frac{p^-}{m^2}\tau \\ \frac{dX^I}{d\tau} &= \frac{p^I}{m^2} \Rightarrow X^I(\tau) = x_0^I + \frac{p^I}{m^2}\tau. \end{aligned} \tag{16.42}$$

Schrödinger Picture

The quantum phase space variables are (X^I, x_0^-, p^I, p^+) (or x_0^I instead of X^I), and we can postulate the canonical commutation relations for them, as usual replacing the Poisson brackets with commutators

$$\begin{aligned} [X^I, p^J] &= i\eta^{IJ} = i\delta^{IJ} \\ [x_0^-, p^+] &= i\eta^{-+} = -i. \end{aligned} \tag{16.43}$$

The other commutators are zero.

Heisenberg Picture

In the Heisenberg picture, the quantum phase space variables turn into time-dependent operators $(X^I(\tau), X^-(\tau), p^I(\tau), p^+(\tau))$, and the canonical commutation relations are

$$\begin{aligned} [X^I(\tau), p^J(\tau)] &= i\eta^{IJ} = i\delta^{IJ} \\ [X^-(\tau), p^+(\tau)] &= i\eta^{-+} = -i\ , \end{aligned} \tag{16.44}$$

and the rest are zero.

Besides these variables, we can define also the dependent variables as operators,

$$X^+(\tau) \equiv \frac{p^+}{m^2}\tau$$
$$p^- = \frac{1}{2p^+}(p^I p^I + m^2). \tag{16.45}$$

In fact, in light cone coordinates, one customarily defines x^+ as light-cone "time", and therefore the generator associated with translations in it,

$$p^- \leftrightarrow \frac{\partial}{\partial X^+}\ , \tag{16.46}$$

as light-cone energy. On the other hand, the Hamiltonian is the generator of τ translations, hence it is

$$\frac{\partial}{\partial \tau} = \frac{p^+}{m^2}\frac{\partial}{\partial X^+} \leftrightarrow \frac{p^+}{m^2}p^-. \tag{16.47}$$

Then we must define the Heisenberg picture Hamiltonian as

$$H(\tau) = \frac{p^+(\tau)}{m^2}p^-(\tau) = \frac{1}{2m^2}(p^I(\tau)p^I(\tau) + m^2) \leftrightarrow \frac{\Box + m^2}{2m^2}. \tag{16.48}$$

We still need to define the state space on which the Heisenberg operators act. A basis of states is the momentum basis $|p^+, \vec{p}_T\rangle \equiv |p^+, p^I\rangle$. On it, the Hamiltonian acts as

$$H|p^+, \vec{p}_T\rangle = \frac{1}{2m^2}(p^I p^I + m^2)|p^+, \vec{p}_T\rangle. \tag{16.49}$$

A general state then is defined by a wavefunction

$$|\psi, \tau\rangle = \int dp^+ d\vec{p}_T \psi(\tau, p^+, \vec{p}_T)|p^+, \vec{p}_T\rangle\ , \tag{16.50}$$

so that the Schrödinger equation is

$$i\frac{\partial}{\partial \tau}\psi(\tau, p^+, \vec{p}_T) = \frac{1}{2m^2}(p^I p^I + m^2)\psi(\tau, p^+, \vec{p}_T). \tag{16.51}$$

Important Concepts to Remember

- The action of a relativistic particle is $S_1 = -m\int d\tau = -m\int d\tau\sqrt{-\dot{X}^2}$.
- The action is invariant under reparametrizations and its equation of motion is the free motion.
- Coupling to background fields results in the motion of the particle interacting with these fields.

- In quantum field theory in the worldline formalism, we define Feynman diagrams ad hoc: the propagator is given from the particle action, the vertex is defined separately, as is the rule for integrating over vertices.
- In field theory, consistency doesn't restrict the form of vertices, hence we have many interacting quantum field theories. In string theory, it is completely constrained, so we will find a unique string theory.
- We can write a first order particle action S_P that is quadratic in fields and derivatives, and has a smooth $m \to 0$ limit.
- The action has reparametrization invariance, which can be fixed by the gauge choice $e = 1$, where $S_P = 1/2 \int d\tau \dot{X}^2$. The Gauss constraint is the vanishing of the energy-momentum tensor $T_{\gamma\gamma} = 0$.
- In light-cone gauge $X^+ = p^+ \tau / m^2$, quantization is in terms of the independent (physical) variables (X^I, X^-, p^I, p^+), and p^- acts as light-cone energy.

Further reading: See [20] (at an advanced undergraduate/beginning graduate student level), [21] (at advanced graduate level) and [22] (at an advanced graduate level; very good for basics, however it doesn't have modern topics) for more on string theory.

Exercises

(1) Calculate the geodesic equation for motion of a free particle in a gravitational field, coming out of the particle action S_1, with $\eta_{\mu\nu} \to g_{\mu\nu}$.
(2) Write down a particle action, second order in derivatives, for motion on a circle.
(3) Quantize covariantly the particle in the gauge $e = 1$.
(4) Show that the equation of motion for the particle action plus the coupling $\int d^4x j^\mu A_\mu$ is the relativistic Lorentz force law. Repeat the derivation in the non-relativistic limit.

Chapter 17
Relativistic Strings

In this chapter we will finally start describing string theory. String theory is a theory of relativistic strings. That means, it is not like violin strings, or cosmic strings, which are essentially non-relativistic and as a result can have compression modes, along the string. Also, unlike them, they represent an idealization, and have zero thickness, i.e., they are truly one-dimensional, and moreover their endpoints will move at the speed of light. But like them, will be characterized by a certain string tension, or energy per unit length, T.

17.1 String Worldsheet and Tension

Note that the idealization is one step better than the idealization of having point particles, which is a very counter-intuitive one when we think about it, yet only a little bit so: in the direction transverse to the string, it is still point-like. Because of this, the string moving in time will span a two-dimensional surface called a *worldsheet*. The surface will be characterized by *intrinsic* coordinates on the worldsheet, like always in curved spaces, σ for worldsheet length (along the string) and τ for worldsheet time, see Fig. 17.1. On the other hand, in spacetime we will have a path depending on (σ, τ), $X^\mu = X^\mu(\sigma, \tau)$, generalizing the case of $X^\mu(\tau)$ for the particle.

Just like the action for the particle was a mass parameter $-m$ times the length of its worldline, $\int d\tau$, the action of a string is expected to be a the mass parameter $-T$ times the area of its worldsheet. In this way, the equation of motion will minimize the area of the worldsheet, just like the equation of motion for the particle minimizes the length of the path, giving the geodesic.

For historical reasons, the string tension is written as $T = 1/(2\pi\alpha')$. Indeed, string theory did not start the way we describe it here, as a theory of strings, but rather with an amplitude (the Veneziano amplitude) that was found to describe hadronic physics somewhat. It was then found that it reproduces a linear relation between spin and mass squared of hadrons, $J = \alpha' M^2 + J_0$, that can be obtained from rotating strings. Then one finds that $T = 1/(2\pi\alpha')$ is the tension of the strings. Therefore the

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_17

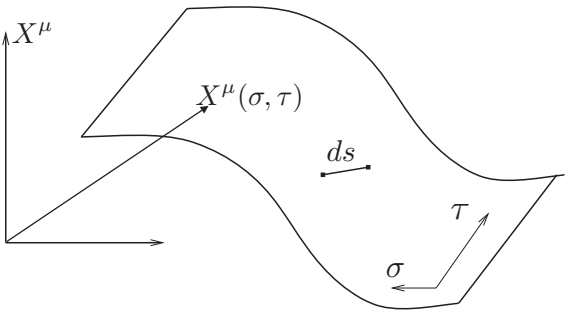

Fig. 17.1 String worldsheet moving in spacetime

action is

$$S = -T \int dA = -\frac{1}{2\pi\alpha'} \int dA. \tag{17.1}$$

We see that α' has dimension of length squared. But as we know from general relativity, the invariant area of the worldsheet is $d^2\xi\sqrt{-\det(\gamma_{ab})}$, where $\{\xi\} = (\sigma, \tau)$ for $a, b = 0, 1$. The action would then be

$$S = -\frac{1}{2\pi\alpha'} \int d\sigma d\tau \sqrt{-\det(\gamma_{ab})} = -\frac{1}{2\pi\alpha'} \int_{\tau_i}^{\tau_f} d\tau \int_0^l d\sigma \sqrt{-\gamma_{00}\gamma_{11} + \gamma_{12}^2}\,, \tag{17.2}$$

but γ_{ab} would be the intrinsic metric on the worldsheet.

17.2 Induced Metric and the Nambu–Goto Action

However, we want that the action defines the embedding of the string worldsheet in spacetime, $X^\mu X^\mu(\xi)$, i.e., the string trajectory.

For simplicity, to understand the embedding, we start with a spatial 2-dimensional surface embedded into an Euclidean space, with metric

$$ds^2 = d\vec{X} \cdot d\vec{X}. \tag{17.3}$$

Consider a surface spanned by $\{\xi_i\}$, $i = 1, 2$, with embedding $\vec{X} = \vec{X}(\xi)$. Then on the embedding, we have

$$d\vec{X} = \frac{\partial \vec{X}}{\partial \xi^i} d\xi^i\,, \tag{17.4}$$

and replacing it inside the Euclidean metric, we obtain the induced line element,

$$ds^2_{\text{induced}} = \left(\frac{\partial \vec{X}}{\partial \xi^i} \cdot \frac{\partial \vec{X}}{\partial \xi^j}\right) d\xi^i d\xi^j \equiv g_{ij}(\xi) d\xi^i d\xi^j\,, \tag{17.5}$$

where the induced metric is

$$g_{ij}(\xi) = \frac{\partial \vec{X}}{\partial \xi^i} \cdot \frac{\partial \vec{X}}{\partial \xi^j}. \tag{17.6}$$

We can now repeat this for a 1+1 dimensional (Minkowski signature) surface, embedded into an arbitrary space of Minkowski signature, with metric $g_{\mu\nu}(X)$, and

$$ds^2 = g_{\mu\nu} dX^\mu dX^\nu. \tag{17.7}$$

As before, the embedding $X^\mu(\xi^a)$ implies that on the embedding we have

$$dX^\mu = \frac{\partial X^\mu}{\partial \xi^a} d\xi^a \ , \tag{17.8}$$

so the line element is

$$ds^2_{\text{induced}} = g_{\mu\nu}(X) \frac{\partial X^\mu}{\partial \xi^a} \frac{\partial X^\nu}{\partial \xi^b} d\xi^a d\xi^b. \tag{17.9}$$

Finally then, the induced metric on the worldsheet, or "pullback" of the spacetime metric is

$$h_{ab}(\xi) = \partial_a X^\mu \partial_b X^\nu g_{\mu\nu}(X). \tag{17.10}$$

We are now ready to write the action for the string in the form found by Nambu and Goto. Note that Yoichiro Nambu got the Nobel prize mostly for the Nambu–Goldstone boson, but the string action was also cited by the Nobel committee, so to date this is the only Nobel prize related to string theory. The Nambu–Goto action is written with the above induced metric instead of γ_{ab}, so

$$S_{NG} = -\frac{1}{4\pi\alpha'} \int d\sigma \, d\tau \sqrt{-\det(h_{ab}(\xi))}. \tag{17.11}$$

Symmetries

The Nambu–Goto action is

- reparametrization (general coordinate, or diffeomorphism) invariant, like the particle action was. The action in terms of the intrinsic metric γ_{ab} is explicitly reparametrization invariant (as we know from general relativity), but so is the action in terms of the induced metric $h_{ab}(\xi)$, since for a transformation $\xi'^a = \xi'^a(\xi^b)$, it transforms as

$$h_{ab}(\xi) = g_{\mu\nu} \frac{\partial X^\mu}{\partial \xi'^c} \frac{\partial X^\nu}{\partial \xi'^d} \frac{\partial \xi'^c}{\partial \xi^a} \frac{\partial \xi'^d}{\partial \xi^b} = h'_{cd}(\xi') \frac{\partial \xi'^c}{\partial \xi^a} \frac{\partial \xi'^d}{\partial \xi^b}. \tag{17.12}$$

That means that (by taking the determinant on both sides)

$$\sqrt{-\det(h_{ab})} = \sqrt{-\det(h'_{ab})}\left|\frac{\partial \xi'}{\partial \xi}\right| \Rightarrow d^2\xi\sqrt{-\det(h_{ab})} = d^2\xi'\sqrt{-\det(h'_{ab})}\,, \tag{17.13}$$

so indeed the action is invariant.

- spacetime Poincaré invariant, in the case that we will study further, of flat spacetime, $g_{\mu\nu}(X) = \eta_{\mu\nu}$.

17.3 The First Order Polyakov Action

Like in the case of the particle, we note that the Nambu–Goto action is highly nonlinear, so it is not ideal for quantization. For that, we must write a first order form that is quadratic in the fields. In fact, the coordinates $X^\mu(\xi)$ act as scalars in the quantum field theory on the two dimensional worldsheet. In order to write a first order action, we must introduce auxiliary fields, and like in the case of the particle, we know that we can use an intrinsic worldsheet metric γ_{ab} for that (though in the particle case, we have written the action instead in terms of the einbein, since it was somewhat easier).

We can easily write the first order action simply by thinking of what we want to obtain. As we said, we will want to write the action in flat spacetime ($g_{\mu\nu} = \eta_{\mu\nu}$). We need a quadratic action for two dimensional massless scalars, with a prefactor $1/(2\pi\alpha')$. The unique result is the *Polyakov action*, (which was first written by Brink, Di Vecchia, Howe, Deser and Zumino...)

$$S_P[X,\gamma] = -\frac{1}{4\pi\alpha'}\int_M d\sigma\, d\sigma\sqrt{-\gamma}\gamma^{ab}\partial_a X^\mu \partial_b X^\nu \eta_{\mu\nu}. \tag{17.14}$$

The action is equivalent to the Nambu–Goto action upon solving for the intrinsic metric γ_{ab}.

Varying S_P with respect to γ^{ab}, we obtain (using $\delta \det \gamma_{ab}/\det\gamma_{ab} = \gamma^{ab}\delta\gamma_{ab} = -\gamma_{ab}\delta\gamma^{ab}$)

$$\delta S_P = -\frac{1}{4\pi\alpha'}\int_M d\sigma\, d\tau\sqrt{-\gamma}\gamma^{ab}\left[\partial_a X^\mu\partial_b X^\nu\eta_{\mu\nu} - \frac{1}{2}\gamma_{ab}(\gamma^{cd}\partial_a X^\mu\partial_b X^\nu\eta_{\mu\nu})\right]. \tag{17.15}$$

The equation of motion of γ^{ab}, $\delta S_P/\delta\gamma^{ab}$=0 is, identifying h_{ab}=$\partial_a X^\mu\partial_b X^\nu\eta_{\mu\nu}$,

$$h_{ab} - \frac{1}{2}\gamma_{ab}(\gamma^{cd}h_{cd}) = 0. \tag{17.16}$$

Multiplying by h^{ab}, we obtain $(\gamma^{ab}h_{ab})^2 = 4$, which is solved by

$$\gamma_{ab} = h_{ab} = \partial_a X^\mu\partial_b X^\nu\eta_{\mu\nu}. \tag{17.17}$$

Then indeed, solving the equation of motion for γ_{ab} and it putting back in S_P, we obtain $S_P = S_{NG}$, hence the Polyakov action is a first order form for Nambu–Goto.

Symmetries

The Polyakov action has the following symmetries:

- spacetime Poincaré invariance, as before.
- worldsheet reparametrization (diffeomorphism) invariance, defined by 2 transformations $\sigma'(\sigma, \tau)$ and $\tau'(\sigma, \tau)$. Under it, the coordinates X^μ are scalars, as we already said, so $X'^\mu(\sigma', \tau') = X^\mu(\sigma, \tau)$, whereas the intrinsic metric transforms in the usual way,
$$\gamma_{ab}(\sigma, \tau) = \gamma'_{cd}(\sigma', \tau') \frac{\partial \xi'^c}{\partial \xi^a} \frac{\partial \xi'^d}{\partial \xi^b}. \tag{17.18}$$
- Weyl invariance. This is an extra symmetry with respect to the Nambu–Goto action, and is also a local invariance, that acts on the metric, but not on the two dimensional coordinates ξ^a, or the scalars X^μ, i.e.,
$$\begin{aligned} X'^\mu(\sigma, \tau) &= X^\mu(\sigma, \tau) \\ \gamma'_{ab}(\sigma, \tau) &= e^{2\omega(\sigma,\tau)} \gamma_{ab}(\sigma, \tau); \quad \forall \omega(\sigma, \tau). \end{aligned} \tag{17.19}$$

Indeed, in two dimensions, we easily see that $\sqrt{-\det \gamma_{ab}} \to e^{2\omega} \sqrt{-\det \gamma_{ab}}$, but $\gamma^{ab} \to e^{-2\omega} \gamma^{ab}$.

This extra Weyl symmetry will be very important in the following, showing that the Polyakov form of the string action is more powerful.

17.4 Equations of Motion, Boundary Conditions and Constraints

We now vary the action, in order to obtain the equations of motion (from the bulk term), boundary conditions (from the boundary term) and constraints (from gauge fixing).

Before considering the complete boundary conditions, we can make an observation about them. The strings have spatial extension, so we need boundary conditions for the endpoints, which then can be *open*, i.e., the they are different points in spacetime, or *closed*, which means that they are at the same points in spacetime.

Before continuing, we note that the energy-momentum tensor (or more precisely, the Belinfante tensor) is obtained by varying the action with respect to the metric. In string theory, one usually considers a different normalization that normal, with an extra factor of 2π, i.e.,
$$T_{ab}(\sigma, \tau) = -\frac{4\pi}{\sqrt{-\gamma}} \frac{\delta S_P}{\delta \gamma^{ab}}. \tag{17.20}$$

In string theory, it is conventional to multiply by this 2π, in order for the result to be simpler, namely

$$T_{ab}(\sigma,\tau) = \frac{1}{\alpha'}\left[\partial_a X^\mu \partial_b X_\mu - \frac{1}{2}\gamma_{ab}\partial_c X^\mu \partial^c X_\mu\right]. \tag{17.21}$$

Note that normally, this would not be possible, since the energy-momentum tensor couples to gravity through Einstein's equation. But in two dimensions, we cannot add a kinetic term for gravity, i.e., an Einstein–Hilbert term on the worldsheet, since this term,

$$\frac{1}{4\pi}\int_M d^2\xi\sqrt{-\gamma}\mathcal{R}^{(2)} = \chi = 2(1-g)\ , \tag{17.22}$$

is a topological invariant, the Euler number, an integer described by $g = 0, 1, 2, \ldots$, the *genus* of the surface M. That means that the variation with respect to the metric gives zero, not the Einstein tensor $G_{ab} = R_{ab} - \frac{1}{2}g_{ab}\mathcal{R}^{(2)}$ as in higher dimensions. In other dimensions, the Einstein equation $\delta S/\delta\gamma^{ab} = 0$ is $G_{ab} = 8\pi G_N T_{ab}$, so rescaling T_{ab} matters, but in two dimensions, the equation of motion is

$$T_{ab} = 0\ , \tag{17.23}$$

for which a rescaling is irrelevant. As usual, the energy-momentum tensor is conserved,

$$\nabla_a T^{ab} = 0. \tag{17.24}$$

Since the action S_P is Weyl invariant, and Weyl invariance corresponds to multiplying all metric components by the same value, Weyl invariance is equivalent to the statement that

$$\gamma^{ab}\frac{\delta S_P}{\delta\gamma^{ab}} = 0 \Rightarrow T^a{}_a = 0\ , \tag{17.25}$$

i.e., that the energy-momentum tensor is *traceless off-shell*. But note that it was essential to have Weyl invariance, otherwise the energy-momentum tensor need not be traceless off-shell.

We now vary the action with respect to $X^\mu(\sigma,\tau)$, for which we need to partially integrate, obtaining

$$\delta_X S_P = \frac{1}{2\pi\alpha'}\int d\tau\int_0^l d\sigma\sqrt{-\gamma}\delta X^\mu\nabla^2 X_\mu - \frac{1}{2\pi\alpha'}\int d\tau\sqrt{-\gamma}\delta X^\mu\ \partial_\sigma X_\mu\big|_{\sigma=0}^{\sigma=l}. \tag{17.26}$$

We have considered here only a boundary in σ, not in τ, for the string worldsheet, leading to the above boundary term.

Boundary Conditions

The vanishing of the boundary term leads to the boundary conditions,

$$\delta X^\mu \, \partial_\sigma X_\mu \big|_{\sigma=0}^{\sigma=l} = 0, \tag{17.27}$$

which can be split as follows:

–*closed strings*: The boundary conditions are periodic,

$$X^\mu(l,\tau) = X^\mu(0,\tau); \quad \gamma_{ab}(l,\tau) = \gamma_{ab}(0,\tau) \ , \tag{17.28}$$

or more generally,

$$X^\mu(\sigma + l,\tau) = X^\mu(\sigma,\tau); \quad \gamma_{ab}(\sigma + l,\tau) = \gamma_{ab}(\sigma,\tau). \tag{17.29}$$

–*open strings*. For them, we can either consider

- Neumann boundary conditions (vanishing of the derivative of the field at the boundary),

$$\partial_\sigma X^\mu(l,\tau) = \partial_\sigma X^\mu(0,\tau) = 0 \ , \tag{17.30}$$

  where the endpoints of the string are free, and as we will shortly see, they move at the speed of light.
- Dirichlet boundary conditions (the field at the boundary is constant),

$$\delta X^\mu(l,\tau) = \delta X^\mu(0,\tau) = 0. \tag{17.31}$$

  In this case, the endpoints of the string are fixed, and we will see in the chapter after next that this corresponds to the strings ending on objects called D-branes. We will therefore not study them further in this chapter.

Equations of Motion

The vanishing of the bulk variation of the action gives the equation of motion, which is

$$\nabla^2 X^\mu = 0 \ , \tag{17.32}$$

the wave equation in 1+1 dimensions.

Gauge Fixing and Constraints

We have seen that the Polyakov action has 3 local worldsheet invariances, i.e., arbitrary functions of (σ,τ):

- two diffeomorphisms, $\sigma'(\sigma,\tau)$ and $\tau'(\sigma,\tau)$
- one Weyl, $\omega(\sigma,\tau)$.

On the other hand, the metric is a symmetric matrix, so has also 3 independent components, γ_{00}, γ_{01} and γ_{11}. That means that we can fix the metric to anything that we want. Of course, one should check that we can indeed reach the domain of values for the components in which the desired value resides, but that is true in the cases of interest.

The simplest choice is usually called "conformal gauge", and is

$$\gamma_{ab} = \eta_{ab}. \tag{17.33}$$

Note that sometimes this is called (more rigorously) "unit gauge", but we will use the term conformal gauge in the following.

In the conformal gauge, the Polyakov action becomes

$$S_P = -\frac{1}{4\pi\alpha'} \int d^2\sigma \eta_{\mu\nu}\eta^{ab}\partial_a X^\mu \partial_b X^\nu \ , \tag{17.34}$$

and correspondingly, the equations of motion become

$$\Box X^\mu = \left(-\frac{\partial^2}{\partial\tau^2} + \frac{\partial^2}{\partial\sigma^2}\right) X^\mu = 0. \tag{17.35}$$

But when we fix a gauge condition, we must impose the equation of motion of the gauge fixed object as a constraint. Since we have fixed γ_{ab}, we must impose its equation of motion,

$$T_{ab} = 0 \ , \tag{17.36}$$

as a constraint.

Static Gauge and the String Tension

Another useful gauge is the static gauge, in which the string looks static, and we fix time X^0 to be equal to the worldsheet time τ,

$$X^0 = \tau; \quad \vec{X} = \vec{X}(\sigma). \tag{17.37}$$

We could have chosen also $X^1 = \sigma$, but we want to keep things a bit more general. Then for a static string $\vec{X} = \vec{X}(\sigma)$, we find

$$S_{NG} = -\frac{1}{2\pi\alpha'} \int d\sigma d\tau \sqrt{(\dot{X}\cdot X')^2 - \dot{X}^2 X'^2} = -\frac{1}{2\pi\alpha'} \int_0^l d\sigma \int d\tau \sqrt{|\vec{X}'|} \ , \tag{17.38}$$

where dot means time ($\tau = X^0$) derivative and prime means σ derivative. Then the action becomes

$$S = -\Delta t T \int d\sigma \frac{|d\vec{X}|}{d\sigma} = -\Delta t |\Delta\vec{X}| T \ , \tag{17.39}$$

and since on a static configuration, the action equals minus the energy times Δt, we find

$$E = |\Delta\vec{X}| T \ , \tag{17.40}$$

i.e., T is energy per unit length, deserving thus the name of string tension.

Open String Endpoints

We finally prove that the open string endpoints, in the case of Neumann (free) boundary conditions, move at the speed of light.

In conformal gauge, $\gamma_{ab} = \eta_{ab}$, the energy-momentum tensor is

$$T_{ab} = \frac{1}{\alpha'}\left[\partial_a X^\mu \partial_b X_\mu - \frac{1}{2}\eta_{ab}\partial_c X^\mu \partial^c X_\mu\right]. \tag{17.41}$$

Using the notation above, we find

$$\alpha' T_{01} = \alpha' T_{10} = \dot{X}\cdot X'; \quad \alpha' T_{00} = \alpha' T_{11} = \frac{1}{2}(\dot{X}^2 + X'^2). \tag{17.42}$$

Since at endpoints, the Neumann boundary condition is $\partial_\sigma X^\mu(\sigma,\tau) \equiv X' = 0$, the conditions become simply

$$0 = \dot{X}^2 = \frac{\partial X^\mu}{\partial\tau}\frac{\partial X^\nu}{\partial\tau}\eta_{\mu\nu}\ , \tag{17.43}$$

which amounts to the fact that the endpoints move on $ds^2 = 0$, i.e., at the speed of light.

17.5 Coupling of String to Background Fields

We have seen how to couple the fields to spacetime gravity in the Polyakov action. We simply need to replace $\eta_{\mu\nu}$ with the general metrix $g_{\mu\nu}(X)$. But we need to maintain the local symmetries of the original action. In fact, in string theory, we start with an action in a particular background, here flat space, and quantize the theory (as we will do in the next chapter) in this background. But then we will see that the massless states of the string are fields $A_{\mu\nu}$, whose symmetric traceless part is identified with the graviton $h_{\mu\nu}$, the antisymmetric part with a tensor $B_{\mu\nu}$, and the trace with a field ϕ called the dilaton. These fields will condense and modify the background, but this means that the modified background should still satisfy the initial symmetries, if there are no anomalies (if the symmetries are preserved at the quantum level, as they should be, for a *local* symmetry). We have then consistency conditions on the possible backgrounds.

With this caveat, we can turn to writing the general Polyakov action in a consistent background of $g_{\mu\nu}$, $B_{\mu\nu}$ and ϕ. Since $B_{\mu\nu}$ is the antisymmetric part of the field whose symmetric traceless part is $g_{\mu\nu}$, it will couple in the action in the same way, via the *pull-back* from spacetime to the worldsheet of $B_{\mu\nu}$,

$$B_{\mu\nu}\partial_a X^\mu \partial_b X^\nu\ , \tag{17.44}$$

and since moreover this object is antisymmetric in (ab), we can make an invariant out of it by multiplying with ϵ^{ab} instead of $\sqrt{-\gamma}\gamma^{ab}$, for an action

$$S_P = -\frac{1}{4\pi\alpha'}\int d^2\sigma \partial_a X^\mu \partial_b X^\nu \left[\sqrt{-\gamma}\gamma^{ab} g_{\mu\nu}(X) + \alpha' \epsilon^{ab} B_{\mu\nu}(X)\right]. \quad (17.45)$$

The α' in the second term is conventional, to make $B_{\mu\nu}$ dimensionful.

Finally, to add a coupling to the dilaton, we must consider a term that would vanish for *constant* ϕ, since ϕ is not necessarily zero. It cannot be a normal scalar, since that would have the interpretation of an extra coordinate X. The only possibility, once we think about it, is to multiply the $\mathcal{R}^{(2)}$, whose integral gives the topological invariant (Euler number), i.e., to add

$$\frac{1}{4\pi}\int d^2\sigma \phi(X)\mathcal{R}^{(2)} \quad (17.46)$$

to S_P.

Important Concepts to Remember

- String theory is a theory of relativistic idealized strings, defined on a worldsheet (σ, τ), with the coordinates appearing as worldsheet scalars $X^\mu(\sigma, \tau)$.
- The Nambu–Goto action is the minus the string tension times the invariant area of the worldsheet, calculated with the induced metric on the worldsheet, or the "pullback" of the spacetime metric.
- The first order form of the Nambu–Goto action is the Polyakov action, with an independent (intrinsic) metric γ_{ab}, and with the form of a kinetic action for the scalars $X^\mu(\sigma, \tau)$.
- The Polyakov action is spacetime Poincaré invariant, diffeomorphism invariant and Weyl invariant.
- The boundary conditions for the string are: closed, open Neumann and open Dirichlet.
- The equations of motion for X^μ are simply the two dimensional wave equation.
- We can fix a gauge for diffeomorphisms and Weyl invariance, fixing the metric γ_{ab} to anything, in particular to the conformal gauge $\gamma_{ab} = \eta_{ab}$.
- In a gauge like the conformal gauge, the metric γ_{ab} equation of motion, $T_{ab} = 0$, is imposed as a constraint.
- The open strings move at the speed of light, and in a static gauge, the string energy per unit length is the string tension.
- We can write the Polyakov action in a background for $g_{\mu\nu}$, $B_{\mu\nu}$, ϕ.

Further reading: See [20–22].

Exercises

(1) Calculate the Hamiltonian and the equations of motion for the Nambu–Goto string in static gauge.
(2) Consider the Polyakov action

$$S_P = -\frac{1}{4\pi\alpha'}\int d\sigma d\tau[\sqrt{-\gamma}\gamma^{ab}\partial_a X^\mu \partial_b X^\nu g_{\mu\nu}(X) + \epsilon^{ab}\partial_a X^\mu \partial_b X^\nu B_{\mu\nu}(X)] \,, \tag{17.47}$$

where $g_{\mu\nu}(X)$ is a metric and $B_{\mu\nu}(X)$ is antisymmetric. Derive the Nambu–Goto action corresponding to it.
(3) Expand the Polyakov action in background fields $g_{\mu\nu}$, $B_{\mu\nu}$, ϕ in terms of gauge covariant objects around a spacetime point, using the tangent space at the point.
(4) Can one add a "cosmological constant" term on the 1+1 dimensional worldsheet of the Nambu–Goto string? Why? Compare with the case of a 2+1 dimensional "worldvolume".

Chapter 18
Light-Cone Gauge Strings and Quantization

In the previous chapter, we have described the classical theory of strings, from the point of view of their worldsheet. Here we will describe the way to quantize them, using a particular gauge called light-cone gauge.

18.1 Light-Cone Gauge

We have seen that the Polyakov action, with diffeomorphism and Weyl invariance, can be gauge fixed to put the metric in the Minkowski form, in the conformal gauge, $\gamma_{ab} = \eta_{ab}$. The action becomes

$$S_P = -\frac{1}{4\pi\alpha'}\int_{-\infty}^{+\infty} d\tau \int_0^l d\sigma \eta^{ab}\partial_a X^\mu \partial_b X^\nu \eta_{ab} \ , \tag{18.1}$$

and the equation of motion is the flat space wave equation,

$$\left(-\frac{\partial^2}{\partial\tau^2} + \frac{\partial^2}{\partial\sigma^2}\right) X^\mu(\sigma,\tau) = 0. \tag{18.2}$$

We have used thus 3 arbitrary functions to fix 3 components, so we might think that the gauge invariance is completely fixed. But we know from the case of electromagnetism that this is not so. For instance, in Lorenz gauge or the $A_0 = 0$ gauge, even though it uses one function to fix one component of the gauge field, we can still have *residual* gauge invariance, where the parameter λ of the transformation is not anymore an arbitrary function, but has a special dependence on coordinates (this allows us to fix both $A_0 = 0$ and $\partial^\mu A_\mu = 0$, for the radiation gauge condition).

Similarly now, we have a residual symmetry, a combination of reparametrization (diffeomorphism, or general coordinate) invariance and Weyl invariance, with a

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_18

specific coordinate dependence, that leaves the gauge fixed action invariant. Defining

$$\sigma^+ \equiv \tau + \sigma; \quad \sigma^- \equiv \tau - \sigma \,, \tag{18.3}$$

a subset of general coordinate transformations that we consider is

$$\begin{aligned} \sigma^+ &\to \tilde{\sigma}^+ = f(\sigma^+) \\ \sigma^- &\to \tilde{\sigma}^- = g(\sigma^-). \end{aligned} \tag{18.4}$$

Invariance *of flat space* under these transformations is called "conformal invariance", and it leads to a modification of the flat space metric by multiplication by a factor,

$$ds^2 = -d\sigma^+ d\sigma^- = -\frac{d\tilde{\sigma}^+}{f'}\frac{d\tilde{\sigma}^-}{g'} = -(f'g')^{-1}(\tilde{\sigma}^+, \tilde{\sigma}^-) d\tilde{\sigma}^+ d\tilde{\sigma}^- . \tag{18.5}$$

In general, conformal invariance is invariance under a subset of general coordinate transformations *of flat space* that multiplies the metric with a conformal factor,

$$ds^2 = d\vec{x}^2 = e^{2\omega(\vec{x}')} d\vec{x}'^2 \,, \tag{18.6}$$

such that the action remains the same action *with the original flat metric*.

The string action in conformal gauge is conformal invariant, since after the conformal transformation we can make a compensating Weyl transformation $\gamma_{ab} \to e^{2\omega}\gamma_{ab}$, under which $\sqrt{-\gamma}\gamma^{ab}$ is invariant in 2 dimensions, that gets rid of the conformal factor $(f'g')^{-1} \equiv e^{-2\omega}$, so we are back at the flat space metric.

After the conformal transformation, we have (defining also $\tilde{\sigma}^\pm \equiv \tilde{\tau} \pm \tilde{\sigma}$),

$$\begin{aligned} \tilde{\tau} &= \frac{1}{2}\left(\tilde{\sigma}^+(\sigma^+) + \tilde{\sigma}^-(\sigma^-)\right) \\ \tilde{\sigma} &= \frac{1}{2}\left(\tilde{\sigma}^+(\sigma^+) - \tilde{\sigma}^-(\sigma^-)\right). \end{aligned} \tag{18.7}$$

and, since now $\partial_\pm = \frac{1}{2}(\partial_\tau \pm \partial_\sigma)$, we obtain that $\tilde{\tau}$ satisfies also the free wave equation,

$$\left(-\frac{\partial^2}{\partial\tau^2} - \frac{\partial^2}{\partial\sigma^2}\right)\tilde{\tau} = -\left(\frac{\partial}{\partial\tau} - \frac{\partial}{\partial\sigma}\right)\left(\frac{\partial}{\partial\tau} + \frac{\partial}{\partial\sigma}\right)\tilde{\tau} = -4\frac{\partial}{\partial\sigma^+}\frac{\partial}{\partial\sigma^-}\tilde{\tau}. \tag{18.8}$$

Defining light-cone coordinates in spacetime by

$$X^\pm = \frac{X^0 \pm X^{D-1}}{\sqrt{2}}, \tag{18.9}$$

we see that $\tilde{\tau}$ satisfies the same equation as X^μ, in particular as X^+ (since linear combinations of X^μ's satisfy the same equation). That means that we can choose a linear relation between them, specifically

$$\tilde{\tau} = \frac{X^+}{p^+} + \text{const.}\,, \tag{18.10}$$

or

$$X^+(\sigma, \tau) = x^+ + p^+\tau. \tag{18.11}$$

This is called the *light-cone gauge* condition, and is a gauge fixing for conformal invariance, besides the conformal gauge that fixed diff and Weyl invariances.

Then the general solution of the wave equation is obtained like the form for $\tilde{\tau}$ was written also, namely as a sum of an arbitrary function of σ^+ and an arbitrary function of σ^-,

$$X^\mu = X^\mu_R(\sigma^-) + X^\mu_L(\sigma^+). \tag{18.12}$$

Here X^μ_R are called "right-moving modes" and X^μ_L are "left-moving modes", referring to the fact that these are wave modes moving one way or the other along the string.

18.2 String Mode Expansions

Closed String Mode Expansion

For a closed string, we have periodicity, $X^\mu(\tau, l) = X^\mu(\tau, 0)$, and we can choose a scale (by constant reparametrizations on the worldsheet) such that $l = 2\pi$. The reason is that we can now parametrize the closed string by (σ being) an angle, and correspondingly we can use the Fourier theorem to decompose $X_L(\tau + \sigma)$ and $X_R(\tau - \sigma)$ in Fourier modes.

The general solution of the equations of motion is then

$$\begin{aligned} X^\mu_R(\tau - \sigma) &= \frac{1}{2}x^\mu + \frac{\alpha'}{2}p^\mu(\tau - \sigma) + \frac{i}{2}\sqrt{2\alpha'}\sum_{n\neq 0}\frac{1}{n}\alpha^\mu_n e^{-in(\tau-\sigma)} \\ X^\mu_L(\tau + \sigma) &= \frac{1}{2}x^\mu + \frac{\alpha'}{2}p^\mu(\tau + \sigma) + \frac{i}{2}\sqrt{2\alpha'}\sum_{n\neq 0}\frac{1}{n}\tilde{\alpha}^\mu_n e^{-in(\tau+\sigma)}. \end{aligned} \tag{18.13}$$

Note that we could have written different constants for the two fields, we chose to consider half of the constant for $X^\mu = X^\mu_L + X^\mu_R$. Also note that the sum over n is for $n \neq 0$, but we can put in the linear term as the $n = 0$ term, defining

$$\alpha^\mu_0 = \sqrt{\frac{\alpha'}{2}}p^\mu = \tilde{\alpha}^\mu_0. \tag{18.14}$$

Then the total X^μ is

$$X^\mu(\sigma,\tau) = x^\mu + \alpha' p^\mu \tau + \frac{i}{2}\sqrt{2\alpha'}\sum_{n\neq 0}\frac{1}{n}[\alpha_n^\mu e^{-in(\tau-\sigma)} + \tilde{\alpha}_n^\mu e^{-in(\tau+\sigma)}]. \quad (18.15)$$

Since X_L^μ, X_R^μ must be real, it follows that x^μ, $p^\mu \in \mathbb{R}$, and moreover

$$\alpha_{-n}^\mu = (\alpha_n^\mu)^\dagger; \quad \tilde{\alpha}_{-n}^\mu = (\tilde{\alpha}_n^\mu)^\dagger. \quad (18.16)$$

Open String Mode Expansion

We next consider the open string with Neumann boundary conditions (Dirichlet boundary conditions will be considered in the next chapter),

$$X'^\mu\big|_{\sigma=0,\pi} = 0. \quad (18.17)$$

Note that we have fixed the length of the open string to be $l = \pi$, again conventionally. Then the most general solution amounts to taking $\alpha_n^\mu = \tilde{\alpha}_\mu^n$ in (18.15), such that there are only $\sin n\sigma$ modes in X', which vanish at $\sigma = 0, \pi$,

$$X^\mu(\sigma,\tau) = x^\mu + 2\alpha' p^\mu \tau + i\sqrt{2\alpha'}\sum_{n\neq 0}\frac{1}{n}\alpha_n^\mu e^{-in\tau}\cos n\sigma. \quad (18.18)$$

Again we can absorb the linear term as the $n = 0$ term in the sum, by defining

$$\alpha_0^\mu = \sqrt{2\alpha'}p^\mu. \quad (18.19)$$

18.3 Constraints and Hamiltonian

We saw that the constraints coming from fixing the conformal gauge are $T_{ab} = 0$, or $\alpha' T_{01} = \alpha' T_{10} = \dot{X}\cdot X' = 0$ and $\alpha' T_{00} = \alpha' T_{11} = \frac{1}{2}(\dot{X}^2 + X'^2) = 0$. In the $\sigma^\pm$ coordinates, they are

$$\begin{aligned}\alpha' T_{++} &= \frac{\alpha'}{2}(T_{00} + T_{01}) = \partial_+ X\cdot\partial_+ X = \dot{X}_L^2\\ \alpha' T_{--} &= \frac{\alpha'}{2}(T_{00} - T_{01}) = \partial_- X\cdot\partial_- X = \dot{X}_R^2\,,\end{aligned} \quad (18.20)$$

where the last form is valid only on-shell.

The action in conformal gauge is

$$S_P = \frac{1}{4\pi\alpha'}\int_{-\infty}^{+\infty} d\sigma(\dot{X}^2 - X'^2)\,, \quad (18.21)$$

leading to the worldsheet momentum

$$P^{\mu}_{\tau} = \frac{\dot{X}^{\mu}}{2\pi\alpha'} \ , \tag{18.22}$$

and Hamiltonian

$$H = \int_0^l d\sigma(\dot{X}\cdot P_{\tau} - \mathcal{L}) = \frac{1}{4\pi\alpha'}\int_0^l d\sigma(\dot{X}^2 + X'^2) = \frac{1}{2\pi}\int_0^l d\sigma T_{00}. \tag{18.23}$$

–For the *on-shell open string*, $l = \pi$, and substituting the expansion of X^{μ} and the orthonormality relations

$$\begin{aligned} \int_0^{\pi} d\sigma \cos n\sigma \cos m\sigma &= \delta_{n+m,0} \\ \int_0^{\pi} d\sigma \sin n\sigma \sin m\sigma &= \delta_{n+m,0} \ , \end{aligned} \tag{18.24}$$

we find the Hamiltonian

$$H = \frac{1}{2}\sum_{n=-\infty}^{+\infty} \alpha^{\mu}_{-n}\alpha^{\mu}_{n}. \tag{18.25}$$

–For the *on-shell closed string*, $l = 2\pi$, and substituting the expansion of X^{μ} and the orthonormality relations for the exponentials, we find the Hamiltonian

$$H = \frac{1}{2}\sum_{n=-\infty}^{+\infty} [\alpha^{\mu}_{-n}\alpha^{\mu}_{n} + \tilde{\alpha}^{\mu}_{-n}\tilde{\alpha}^{\mu}_{n}]. \tag{18.26}$$

Constraint Modes

The *Virasoro constraints* $T_{++} = 0$ and $T_{--} = 0$ are still functions on the circle (for the closed string) or on the interval (for the open string), and as such can be expanded into Fourier modes.

–*closed string*. In the closed string case, the Fourier modes are defined as

$$\begin{aligned} L_m &= \frac{1}{2\pi}\int_0^{2\pi} d\sigma e^{-im\sigma} T_{--} = \frac{1}{2\pi\alpha'}\int_0^{2\pi} d\sigma e^{-im\sigma}\dot{X}^2_R = \frac{1}{2}\sum_{n=-\infty}^{+\infty} \alpha^{\mu}_{m-n}\alpha^{\mu}_{n} \\ \tilde{L}_m &= \frac{1}{2\pi}\int_0^{2\pi} d\sigma e^{-im\sigma} T_{++} = \frac{1}{2\pi\alpha'}\int_0^{2\pi} d\sigma e^{-im\sigma}\dot{X}^2_L = \frac{1}{2}\sum_{n=-\infty}^{+\infty} \tilde{\alpha}^{\mu}_{m-n}\tilde{\alpha}^{\mu}_{n} \ , \end{aligned} \tag{18.27}$$

where after the definition, we have used the on-shell form. Note that the factor of $1/(2\pi)$ is due to the conventional factor in the energy-momentum tensor that was added in front.

–*open string*. In the open string case, the Fourier modes are defined as

$$L_m = \frac{1}{2\pi}\int_0^{2\pi} d\sigma(e^{im\sigma}T_{++} + e^{-im\sigma}T_{--}) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} d\sigma e^{im\sigma}T_{++}$$
$$= \frac{1}{8\pi\alpha'}\int_{-\pi}^{+\pi} d\sigma e^{im\sigma}(\dot{X}+X')^2 = \frac{1}{2}\sum_{n=-\infty}^{+\infty}\alpha^\mu_{m-n}\alpha^\mu_m. \tag{18.28}$$

We note then that the Hamiltonian can be written as

$$\begin{aligned} H &= L_0 \ \text{ open} \\ &= L_0 + \tilde{L}_0 \ \text{ closed} \end{aligned} \tag{18.29}$$

Then for the open string, the $H = L_0 = 0$ constraint translates into (writing the sum for $n \geq 1$ and for $n \leq -1$ as twice the sum for $n \geq 1$, and separating the $n = 0$ term)

$$M^2 \equiv -p^\mu p_\mu = -\frac{\alpha_0^2}{2\alpha'} = \frac{1}{\alpha'}\sum_{n\geq 1}\alpha^\mu_{-n}\alpha^\mu_n\,, \tag{18.30}$$

and similarly in the closed string case, the $H = L_0 + \tilde{L}_0 = 0$ constraint becomes

$$M^2 \equiv -p^\mu p_\mu = -\frac{\alpha_0^2 + \tilde{\alpha}_0^2}{\alpha'^2} = \frac{2}{\alpha'}\sum_{n\geq 1}(\alpha^\mu_{-n}\alpha^\mu_n + \tilde{\alpha}^\mu_{-n}\tilde{\alpha}^\mu_n). \tag{18.31}$$

Light-Cone Gauge

Until now, we had considered only the conformal gauge, but now we also impose the fact that we are in the light-cone gauge

$$X^+(\sigma,\tau) = x^+ + p^+\tau. \tag{18.32}$$

By comparing with the general expansion (18.15) and the equivalent one for the open string, we see that we must have

$$\alpha_n^+ = 0, \quad \text{for} \quad n \neq 0. \tag{18.33}$$

We also obtain then $\dot{X}^+ \pm X'^+ = p^+$.

Then the Virasoro constraints $(\dot{X} \pm X')^2 = 0$ become, in spacetime light-cone coordinates (where $v \cdot w = -v^- w^+ - v^+ w^- + v^i w^i$) and in the light-cone gauge become

$$(\dot{X}^- \pm X'^-) = \frac{(\dot{X}^i \pm X'^i)^2}{2(\dot{X}^+ \pm X'^+)} = \frac{(\dot{X}^i \pm X'^i)^2}{2p^+}. \tag{18.34}$$

It is easy to see that in components we obtain

$$\alpha_n^- = \frac{1}{2p^+} \sum_{m\in\mathbb{Z}} \alpha^i_{n-m}\alpha^i_n. \tag{18.35}$$

That means that the light-cone gauge puts $\alpha_n^+ = 0$ and fixes (through the Virasoro constraints) α_n^- from α_n^i. The only independent oscillators are now α_n^i. Indeed, then also the mass formula becomes (for the open string)

$$M^2 = 2p^+p^- - p^i p^i = \frac{1}{\alpha'} \sum_{n\geq 1} \alpha^i_{-n}\alpha^i_n \ , \tag{18.36}$$

and a similar sum only over the transverse indices i for the closed string.

18.4 Quantization of String Modes

We next move to describing quantization. We proceed first with general issues, before focusing on the light-cone gauge quantization. Before the light-cone gauge, the Polyakov action is (18.21), with the canonical momentum $P^\mu_\tau = \dot{X}^\mu/(2\pi\alpha')$.

We can calculate the equal time Poisson brackets, and find

$$\begin{aligned}
&[P^\mu(\sigma,\tau), X^\nu(\sigma',\tau)]_{P.B.} = -\delta(\sigma-\sigma')\eta^{\mu\nu} \\
&[X^\mu(\sigma,\tau), X^\nu(\sigma',\tau)]_{P.B.} = [P^\mu(\sigma,\tau), P^\nu(\sigma',\tau)]_{P.B.} = 0.
\end{aligned} \tag{18.37}$$

From them, we can deduce the Poisson brackets for the modes of the string,

$$\begin{aligned}
&[\alpha^\mu_m, \alpha^\nu_n]_{P.B.} = [\tilde{\alpha}^\mu_m, \tilde{\alpha}^\nu_n]_{P.B.} = -im\delta_{m+n,0}\eta^{\mu\nu} \\
&[p^\mu, x^\nu]_{P.B.} = \eta^{\mu\nu} \\
&[\alpha^\mu_m, \tilde{\alpha}^\nu_n]_{P.B.} = 0.
\end{aligned} \tag{18.38}$$

These however need to be supplemented with the Virasoro constraints $(\dot{X} \pm X')^2 = 0$ ($T_{++} = T_{--} = 0$), or their modes $L_n = 0$ and $\tilde{L}_n = 0$.

Then quantization proceeds as usual, by replacing the Poisson brackets $[,]_{P.B.}$ with the commutator $-i[,]$.

The canonical (equal time) commutation relations are

$$\begin{aligned}
&[P^\mu(\sigma,\tau), X^\nu(\sigma',\tau)] = -i\delta(\sigma-\sigma')\eta^{\mu\nu} \\
&[X^\mu(\sigma,\tau), X^\nu(\sigma',\tau)] = [P^\mu(\sigma,\tau), P^\nu(\sigma',\tau)] = 0.
\end{aligned} \tag{18.39}$$

From them, one finds the canonical commutation relations for the modes α_m^μ, $\tilde{\alpha}_m^\nu$, x^μ, p^μ,

$$\begin{aligned}
[\alpha_m^\mu, \alpha_n^\nu] &= [\tilde{\alpha}_m^\mu, \tilde{\alpha}_n^\nu] = m\delta_{m+n,0}\eta^{\mu\nu} \\
[p^\mu, x^\nu] &= \eta^{\mu\nu} \\
[\alpha_m^\mu, \tilde{\alpha}_n^\nu] &= 0.
\end{aligned} \tag{18.40}$$

We see then that the string modes are usual oscillators with a rescaling,

$$\alpha_m^\mu = \sqrt{m}a_m^\mu; \quad \alpha_{-m}^\mu = \sqrt{m}a_m^{\dagger\mu}; \quad m > 0. \tag{18.41}$$

But we note that as it stands, we have too many modes, as seen from the fact that the timelike modes have negative norm (their commutator has the wrong term). As in electromagnetism, the point is that we still need to impose the constraints, now $L_m = 0$ and $\tilde{L}_m = 0$. In a covariant type of quantization, like the Gupta–Bleuler of electromagnetism, we need to impose the constraints as operators acting on physical states (for the covariant quantization of electromagnetism, Gupta–Bleuler type, one imposes $\partial^\mu A_\mu^{(+)}|\psi\rangle = 0$, the positive frequency part of the operatorial Lorenz condition on physical states), so one would need to impose $L_m|\psi\rangle = \tilde{L}_m|\psi\rangle = 0$.

18.5 Light-Cone Gauge Quantization

But instead, I will continue with the quantization in the light-cone gauge. We saw that there, at the classical level, the Virasoro constraints were solved by writing α_n^- in terms of α_n^i, and we also have $\alpha_n^+ = 0$. At the quantum level we must care about the ordering of the oscillators α_n^i, since they don't commute. The correct solution at the quantum level is written in terms of normal ordering (defined as usual for a and $a^\dagger$'s, by putting the $a^\dagger$'s to the left and a's to the right) as

$$\alpha_n^- = \frac{1}{p^+}\left[\frac{1}{2}\sum_{i=1}^{D-2}\sum_{m\in\mathbb{Z}} :\alpha_{n-m}^i\alpha_m^i: -a\delta_{n,0}\right], \tag{18.42}$$

since only the $n = 0$ terms don't commute. As before, we have $\alpha_m^+ = 0$.

In this way, X^+ and X^- are eliminated completely, and there are no constraints left. That means that we can quantize in a physical gauge, meaning only the physical, independent oscillators α_n^i obey canonical commutation relations.

Open String Spectrum

For $n = 0$, the Virasoro constraint mode $L_n = 0$ gives now, considering carefully the order of oscillators,

$$M^2 \equiv 2p^+p^- - p^i p^i = \frac{1}{\alpha'}(N-a)\ , \tag{18.43}$$

where N is a kind of number operator,

$$N = \sum_{n\geq 1} \alpha^i_{-n}\alpha^i_n = \sum_{n\geq 1} n a^{\dagger i}_n a^i_n\ , \tag{18.44}$$

and n is called the *level*, and a is an ordering constant, calculated as follows. In fact, in L_0, we have the terms (rewriting the $n \geq 1$ and $n \leq -1$ through flipping n)

$$\sum_i \sum_{n\geq 1} n\frac{a^{\dagger i}_n a^i_n + a^i_n a^{\dagger i}_n}{2} = N + \sum_{i=1}^{D-2}\sum_{n\geq 1}\frac{n}{2} = N + \frac{D-2}{2}\sum_{n\geq 1} 1. \tag{18.45}$$

But the infinite sum over n is regularized using the usual zeta function regularization of quantum field theory. The Riemann zeta function,

$$\zeta(s) = \sum_{n\geq 1}\frac{1}{n^s}\ , \tag{18.46}$$

is actually a well-defined function on the complex plane, minus the negative integers, but can be analytically continued to a function defined over the whole complex plane, and then one finds that at $s = -1$ one has

$$\zeta(-1) = \sum_{n\geq 1} n = -\frac{1}{12}. \tag{18.47}$$

Replacing above, we find the normal ordering constant

$$a = \frac{D-2}{24}. \tag{18.48}$$

We see from the mass-shell relation $M^2 = (N-a)/\alpha'$ that string theory is a theory of an infinite number of different kinds of particles, labelled by the level n, somewhat similar to the case of a solid, with phonons with Hamiltonian $H = \sum_n \omega_n N_n$. Just that in the solid case, the different kinds of phonons, with different ω_n's, are not considered really as different kinds of particles, but here it is so, since we are in flat space.

But then we notice that for $N = 1$, more precisely for states $\alpha^i_{-1}|0\rangle = a^{\dagger i}_1|0\rangle$, which have eigenvalue 1 for N, the mass is $M^2 = 1 - a$, yet there are $i = 1, \ldots, D-2$ of them, like for a massless vector (which has only transverse components, unlike a massive vector). That means that we must have $M^2 = 0$ in this case, or

$$a = 1. \tag{18.49}$$

That then implies

$$D = 26 \,, \tag{18.50}$$

so our (bosonic) string must live in 26 dimensions. We will see later that a supersymmetric version of the string must instead live in 10 dimensions, by a similar argument.

We have also seen in the particle case that in light-cone coordinates, x^+ plays the role of (light-cone) time, and therefore its canonical conjugate, p^-, is the Hamiltonian. The light-cone Hamiltonian for the string is then

$$H_{l.c.} = p^- = \frac{p^i p^i}{2p^+} + \frac{1}{2\alpha' p^+}(N-1). \tag{18.51}$$

The vacuum of the open string is the one annihilated by all the oscillators (like for the phonon case), but since we have also p^μ among the quantum operators, the vacuum must also come with an eigenvalue for it, i.e., with momentum.

Therefore the vacuum state is $|0,\vec{k}\rangle$, and it satisfies

$$\begin{aligned} p^+|0,\vec{k}\rangle &= k^+|0,\vec{k}\rangle \\ p^i|0,\vec{k}\rangle &= k^i|0,\vec{k}\rangle \\ \alpha_m^i|0,\vec{k}\rangle &= 0, \end{aligned} \tag{18.52}$$

A general state is then

$$|N,\vec{k}\rangle = \left[\prod_{i=1}^{D-2}\prod_{n\geq 1}\frac{(a_n^{\dagger i})^{N_{in}}}{\sqrt{N_{i_n}!}}\right]|0,\vec{k}\rangle = \left[\prod_{i=1}^{D-2}\prod_{n\geq 1}\frac{(\alpha_n^{\dagger i})^{N_{in}}}{\sqrt{n^{N_{in}}N_{i_n}!}}\right]|0,\vec{k}\rangle. \tag{18.53}$$

We note that the vacuum is a spacetime scalar $|0,\vec{k}\rangle$, with $M^2 = -1/\alpha'$, i.e., it is a *tachyon*. This simply says that the potential for this scalar ϕ has negative curvature at zero, $V''(\phi = 0) < 0$, i.e. the vacuum is *unstable*. The bosonic string is therefore actually unstable around the usual vacuum. But there could be some nonperturbative vacuum at some $\phi_0 \neq 0$, which is a true minimum, and around which bosonic string theory is stable. This is not known until now, despite various tentatives at understanding it.

As we argued before, the next state, $a_1^{\dagger i}|0\rangle$, is a vector state corresponding to some gauge field A_μ, which is massless.

Closed String Spectrum

For the closed string, there are two constraints $L_0 - a = \tilde{L}_0 - a = 0$, and it turns out that the constant a is the same $a = 1$ as before (by a similar argument). Then we can write the M^2 formula in two ways,

$$M^2 = \frac{4}{\alpha'}\sum_{n\geq 1}(\alpha^i_{-n}\alpha^i_n - 1) = \frac{4}{\alpha'}\sum_{n\geq 1}\tilde{\alpha}^i_{-n}\tilde{\alpha}^i_n - 1) \ , \tag{18.54}$$

but we must supplement it with the *level-matching condition* (matching the levels on the left and right sectors) $L_0 = \tilde{L}_0$, which means

$$\sum_{n\geq 1}\alpha^i_{-n}\alpha^i_n = \sum_{n\geq 1}\tilde{\alpha}^i_{-n}\tilde{\alpha}^i_n. \tag{18.55}$$

That means that we can write the mass-shell condition also, symmetrically, as

$$M^2 = \frac{2}{\alpha'}\sum_{n\geq 1}(\alpha^i_{-n}\alpha^i_n + \tilde{\alpha}^i_{-n}\tilde{\alpha}^i_n) = \frac{2}{\alpha'}(N + \tilde{N} - 2) \ , \tag{18.56}$$

but supplemented with the level matching condition

$$N = \tilde{N}. \tag{18.57}$$

In this case the vacuum is still a tachyon $|0, \vec{k}\rangle$, but now with $M^2 = -4/\alpha'$. On the other hand, because of the level-matching condition, the first excited states is the one with $N = \tilde{N} = 1$, which is therefore again massless ($M^2 = 0$). This is a state

$$a_1^{\dagger i}\tilde{a}_1^{\dagger j}|0, \vec{k}\rangle = \alpha^i_{-1}\tilde{\alpha}^j_{-1}|0, \vec{k}\rangle \equiv |ij\rangle. \tag{18.58}$$

So this is a state with two transverse indices $|ij\rangle$, $i, j = 1, \ldots, D-2$. In terms of irreducible representations of the Lorentz ($SO(D)$) group, this splits into the symmetric traceless state ((ij)), the antisymmetric state $[ij]$ and the trace ii.

String theory started as an attempt to describe hadronic (strong interaction) physics via the Veneziano amplitude. Only later it was realized that this amplitude comes from strings, and then it was found that closed strings have this massless symmetric traceless (spin 2) state in the spectrum. It was however known that there are no hadronic states that have spin 2 and are massless, so for a while the interest in string theory waned (also since QCD came along as a very good theory of strong interactions), until it was realized (by John Schwarz) that this simply means that *string theory is a theory of quantum gravity* (since it includes a graviton in its spectrum), and is not a theory of strong interactions.

Now we can therefore say that the massless states of the closed bosonic string are the graviton $g_{\mu\nu}$, an antisymmetric tensor $B_{\mu\nu}$ and a scalar ϕ called the dilaton.

We can then self-consistently consider that these massless modes of the string condense and create a background for the string, like we did in the previous chapter. But we need the consistency condition that when we do so, the symmetries of the action are preserved at the quantum level by the introduction of this background. And we saw in particular that the Polyakov action had conformal invariance. Imposing conformal invariance at the quantum level for a general background gives some consistency conditions that imply equations of motion for the spacetime fields.

Thus, even though we are doing worldsheet string physics, we are in fact able to obtain spacetime information, just by imposing quantum consistency of the theory in a general background for the string. We will however not show here how to do that, since it is a bit complicated.

Important Concepts to Remember

- The Polyakov action in light-cone gauge has conformal invariance, which is invariance under a coordinate transformation that acts on the flat space metric by multiplication by a common factor.
- The light-cone gauge is $X^+ = x^+ + p^+\tau$ and it fixes the longitudinal modes, leaving only transverse modes.
- The constraints $T_{++} = T_{--} = 0$ give the infinite number of Virasoro modes $L_n = \sum_m \alpha^\mu_{n-m}\alpha^\mu_m = 0$ and $\tilde{L}_n = \sum_m \tilde{\alpha}^\mu_{n-m}\tilde{\alpha}^\mu_{n-m} = 0$.
- For the on-shell closed string in conformal gauge, the Hamiltonian is $H = \frac{1}{2}\sum_n(\alpha^\mu_{-n}\alpha^\mu_n + \tilde{a}^\mu_{-n}\tilde{a}^\mu_n)$, and for the on-shell open string we have only the modes without tilde.
- Under quantization, the open string mass-shell condition is $M^2 = 2p^+p^- - \vec{p}^2 = \frac{1}{\alpha'}(N - a)$ and the closed string one is $M^2 = 2p^+p^- - \vec{p}^2 = \frac{2}{\alpha'}(N + \tilde{N} - 2a)$, supplemented with the level-matching condition $N = \tilde{N}$.
- The light-cone gauge Hamiltonian is $H = p^+$ defined by the mass-shell condition.
- From zeta function regularization of the sum of zero point energies, we obtain $a = (D - 2)/24$, and from the consistency of the massless spectrum we obtain $D = 26$.
- The strings have a tachyonic (unstable) ground state, and the closed massless modes are a graviton, antisymmetric tensor and dilaton, meaning that string theory is a theory of quantum gravity unified with other interactions.

Further reading: See [20, 21] for more details.

Exercises

(1) Write down the most general solution for the open string $X^\mu(\sigma, \tau)$ with one Neumann (free) boundary condition and one Dirichlet (fixed) boundary condition.
(2) Write down the states of the first 2 *massive* levels of the bosonic open string and closed string.
(3) Calculate the algebra satisfied classically by the L_m, $\tilde{L}_m$, via Poisson brackets.
(4) Construct a coherent state of the graviton (symmetric traceless massless mode of the closed string), and show that it can be thought of as a gravitational background.
(5) Write the (naive) generator of the Lorentz symmetry $J_{\mu\nu}$ in terms of the oscillators of the string α^μ_n.

Chapter 19
D-Branes and Gauge Fields

In this chapter we will explore the consequences of having Dirichlet boundary conditions for open strings. It will imply the existence of objects called D-branes, that have gauge fields on their "worldvolume".

The idea of having Dirichlet boundary conditions arose early on, through the argument we have, of vanishing the boundary term in the variation of the string action. However, initially it was thought it was a bad idea, since a Dirichlet boundary condition meant that the endpoints of strings are fixed at one point in spacetime, which would break translational invariance, one of the most obvious symmetries of quantum field theory. For a while, the idea of having Dirichlet boundary conditions in *all* space and time directions was considered, what is now called a D-instanton, since this was clearly a type of soliton, and very like a usual instanton.

But then it came to be that other objects were considered. If we have Dirichlet boundary conditions in $D - p - 1$ space dimensions and Neumann boundary conditions in $p + 1$ spacetime dimensions, the open string ends on a $p + 1$ dimensional "wall", a p-brane (a term coined from a generalization of the word membrane), or more precisely a Dp-brane (the D standing in for Dirichlet). More precisely, each endpoint of the string can be on the same, or on different Dp-branes. The wall will be called a "worldvolume", generalizing the worldline of the particle and the worldsheet of the string. Note that really, we are assuming that both endpoints of the string have the same type of boundary conditions for the same coordinates, i.e., that we have NN or DD boundary conditions, though in principle it is possible to have ND coordinates as well (and then, the Dp-branes on which the endpoints of the string lie are of different types).

We will use light-cone coordinates, with worldvolume coordinates $\{X^+, X^-, X^a\}$ for the $p + 1$ NN coordinates parallel to the D-brane, and $\{X^i\}$ for the DD coordinates, transverse to the D-brane, therefore we split the lightcone transverse coordinates as $X^I = (X^a, X^i)$. From the point of view of the D-brane, the transverse coordinates act as scalar fields on the worldvolume, in the same way we saw this happening for the particle or the string.

The reason that Dp-branes were finally considered as interesting, is that in a seminal paper in 1989, Dai, Leigh and Polchinski proved that the Dirichlet boundary

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_19

conditions are not simply inert, but rather that this "wall" is a dynamical object, somewhat like a soliton, the D-brane. The paper proved (through a calculation in string theory) that spacetime closed string modes can collide with this "wall" and excite open string modes on it, including the transverse coordinates, thus making it vibrate.

19.1 The D-Brane Action

If the transverse coordinates are also scalars on the worldvolume, the action for the D-brane must start with an action minimizing the volume of the D-brane, just like the action for the string minimizes the surface of the worldsheet, so with

$$S = -T_p \int d^{p+1}\xi \sqrt{-\det(h_{ab})}\ , \tag{19.1}$$

where as there, h_{ab} is the induced metric on the worldvolume, or "pull-back" of the spacetime metric,

$$h_{ab} = \frac{\partial X^\mu}{\partial \xi^a}\frac{\partial X^\nu}{\partial \xi^b} g_{\mu\nu}(X). \tag{19.2}$$

Here T_p is a D-brane tension, generalizing the string tension T, as energy per unit volume of the worldvolume.

Moreover, also like in the case of the string, we have seen that in string theory $g_{\mu\nu}$ and $B_{\mu\nu}$ are the symmetric and antisymmetric parts of the same field $\alpha^i_{-1}\tilde{\alpha}^j_{-1}|0,\vec{k}\rangle$, so actually the action must have them in the combination $g_{\mu\nu}(X) + \alpha' B_{\mu\nu}(X)$. That means that the action for coupling of X^μ to $g_{\mu\nu}(X)$ and $B_{\mu\nu}(X)$ is

$$S = -T_p \int d^{p+1}\xi \sqrt{-\det\left(\frac{\partial X^\mu}{\partial \xi^a}\frac{\partial X^\nu}{\partial \xi^b}[g_{\mu\nu}(X) + \alpha' B_{\mu\nu}(X)]\right)}. \tag{19.3}$$

Like the Nambu–Goto action for the string, the action has diffeomorphism (general coordinate, or reparametrization) invariance, which can be fixed using a gauge. In the case of the string, we have suggested the use of a *static gauge*, which in this more general case of D-branes can be made more precise. The static gauge means identifying the worldvolume coordinates with $p+1$ of the spacetime coordinates, so

$$X^a = \xi^a; \quad a = 0, 1, \ldots, p\ , \tag{19.4}$$

whereas the transverse coordinates are identified with scalar fields $\phi^a(\xi)$ on the worldvolume through a constant rescaling,

$$X^i(\xi^a) = \frac{\phi^i(\xi^a)}{\sqrt{T_p}}. \tag{19.5}$$

We also expand the spacetime metric in a graviton perturbation,

$$g_{\mu\nu}(X) = \eta_{\mu\nu} + 2\kappa_N h_{\mu\nu}(X). \tag{19.6}$$

Firstly, we consider the case of only the scalars being nontrivial, and $h_{\mu\nu} = B_{\mu\nu} = 0$. Then the action is

$$S = -T_p \int d^{p+1}\xi \sqrt{-\det\left(\eta_{ab} + \frac{\partial_a \phi^i \partial_b \phi^i}{T_p}\right)}, \tag{19.7}$$

and expanding the determinant we obtain

$$\begin{aligned}
&-\det\left(\eta_{ab} + \frac{\partial_a \phi^i \partial_b \phi^i}{T_p}\right) \\
&= -\frac{1}{(p+1)!}\epsilon^{a_1\dots a_{p+1}}\epsilon^{b_1\dots b_{p+1}}(\eta_{a_1 b_1} + \partial_{a_1}\phi^i \partial_{b_1}\phi^i)\dots(\eta_{a_{p+1}b_{p+1}} + \partial_{a_{p+1}}\phi^i \partial_{b_{p+1}}\phi^i) \\
&= -\frac{1}{(p+1)!}\epsilon^{a_1\dots a_{p+1}}\epsilon_{a_1\dots a_{p+1}} - \frac{(p+1)}{(p+1)!}\epsilon^{a_1\dots a_p a_{p+1}}\epsilon_{a_1\dots a_p b_{p+1}}\frac{\partial_{a_{p+1}}\phi^i \partial^{b_{p+1}}\phi^i}{T_p} + \cdots \\
&= 1 + \frac{\partial_a \phi^i \partial^a \phi^i}{T_p} + \cdots
\end{aligned} \tag{19.8}$$

Here in the first equality we wrote the definition of the determinant, in the second equality we dropped the terms containing more than one factor of $\partial_a \phi^i \partial_b \phi^i$ (if there is a single ϕ, since $\partial_{[a}\phi\partial_{b]}\phi = 0$, there are no such higher order terms), and in the third we used the equality (easily proved in components)

$$\epsilon^{a_1\dots a_k a_{k+1}\dots a_n}\epsilon_{a_1\dots a_k b_{k+1}\dots b_n} = -k!(n-k)!\delta^{[a_{k+1}\dots a_n]}_{[b_{k+1}\dots b_n]}. \tag{19.9}$$

Therefore finally, expanding the action in the scalar fields we get

$$\begin{aligned}
S &= -T_p \int d^{p+1}\xi \sqrt{1 + \frac{\partial_a \vec{\phi} \cdot \partial^a \vec{\phi}}{T_p} + \cdots} \\
&\simeq -\int d^{p+1}\xi \left[T_p + \frac{\partial_a \vec{\phi} \cdot \partial^a \vec{\phi}}{2} + \cdots\right].
\end{aligned} \tag{19.10}$$

Next, considering the action with a graviton $h_{\mu\nu}$ in it, and expanding the determinant to first order in the graviton and first order in the scalars ϕ^i,

$$
\begin{aligned}
&- \det\left((\eta_{\mu\nu} + 2\kappa_N h_{\mu\nu})(\delta^\mu_a + \partial_a \phi^\mu)(\delta^\nu_b + \partial_b \phi^\nu)\right) \\
&= - \det\left(\eta_{ab} + 2\kappa_N h_{ab} + 4\kappa_N h_{ai} \frac{\partial_b \phi^i}{\sqrt{T_p}} + \cdots\right) \\
&= 1 + h^a{}_a + 4\kappa_N h_{ai} \frac{\partial^a \phi^i}{\sqrt{T_p}} + \cdots ,
\end{aligned} \tag{19.11}
$$

and ignoring the graviton trace $h^a{}_a$ (supposing it is zero), we find for the action (and keeping the leading kinetic term as well)

$$
S \simeq - \int d^{p+1}\xi \left[T_p + \frac{\partial_a \vec{\phi} \cdot \partial^a \vec{\phi}}{2} + 2\kappa_N \sqrt{T_p} h_{ai} \partial^a \phi^i + \cdots \right]. \tag{19.12}
$$

From it, we find the vertex for ϕ^j with h^a_i,

$$
V_i^{aj} = -2\kappa_N \sqrt{T_p} i k^a \delta_i^j . \tag{19.13}
$$

This vertex represents a coupling of a graviton h_{ai}, which is a closed string mode (living in the bulk of the spacetime) to ϕ^j, which is an open string mode, living on the worldvolume of the D-brane (since the endpoints of the open string are restricted to the D-brane, the open string mode lives on it). Thus this vertex represents exactly what we discussed before, the fact that a closed string can collide with the D-brane and excite an open string mode, thus making it vibrate (indeed, ϕ^i corresponds to a fluctuation in the transverse position of the D-brane), as in Fig. 19.1.

By matching with a string theory calculation, coming from a worldsheet with the topology of a disk, that ends on the D-brane, and contains a very thin tube going to infinity (as the closed string mode coming from infinity to collide with the D-brane), we find the same vertex, and we can calculate the tension as a function of string theory objects,

$$
T_p = \frac{1}{(2\pi\alpha')^2 g_{p+1}^2} , \tag{19.14}
$$

where g_{p+1}, the coupling of the $p+1$ dimensional theory on the Dp-brane, is

$$
g_{p+1}^2 = (2\pi)^{p-2} g_s \alpha'^{\frac{p-3}{2}} . \tag{19.15}
$$

We still need to consider the coupling of the D-brane with the last massless mode of the closed bosonic string, the dilaton ϕ, which appears as the trace of the same mode $\alpha^i_{-1}\tilde{\alpha}^i_{-1}|0,\vec{k}\rangle$ as for the graviton and B-field. But the dilaton gives the string coupling g_s through its VEV, by $e^{\langle\phi\rangle} = g_s$. Indeed, string theory has no dimensionless parameter (and only one dimensional parameter, α'), so the only dimensionless parameters come from the vacuum. An action for closed string modes then would be

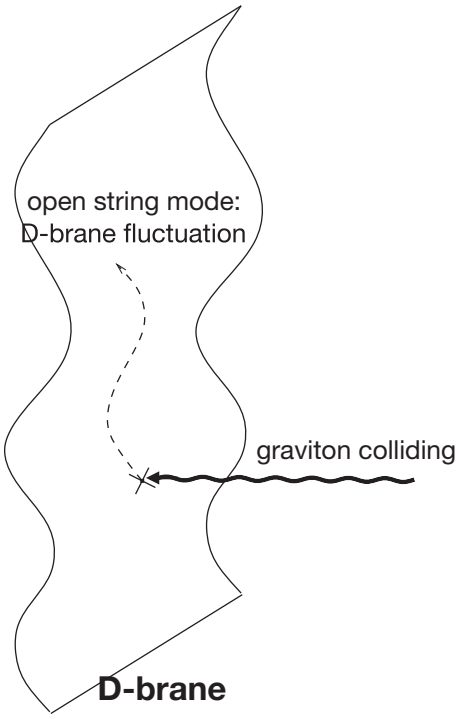

Fig. 19.1 Graviton collides with D-brane and excites the open string mode living on it, making it vibrate

expected to have $1/g_s^2$ in front, and an action for open string modes, like the D-brane action is, to have $1/g_o^2$, g_o being the open string coupling constant.

We can see pictorially that a closed string interaction is the square of an open string interaction, (closed string) = (open string)^2. Indeed, if we cut a basic closed string interaction, a "pair of pants", for a closed string splitting into two, we obtain two identical open string interactions, as in Fig. 19.2, which means that

$$g_s = g_o^2. \tag{19.16}$$

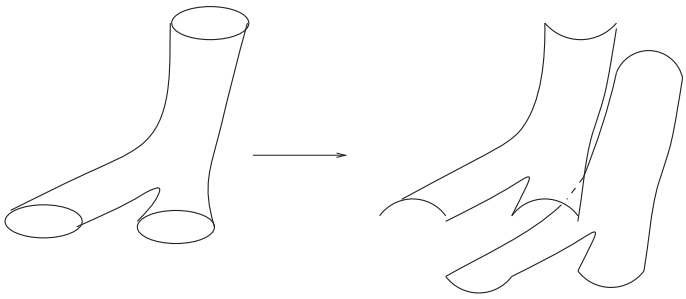

Fig. 19.2 Cutting a basic closed string interaction, a "pair of pants", into two pieces, we obtain two basic open string interactions

That means that we expect to find in front of the D-brane action a factor $\frac{1}{g_o^2} = \frac{1}{g_s} = e^{-\langle\phi\rangle}$. But it follows logically that the full dependence of the D-brane action on ϕ must be via a factor of $e^{-\phi}$, as

$$S_p = -T_p \int d^{p+1}\xi e^{-\phi} \sqrt{-\det\left(\frac{\partial X^\mu}{\partial \xi^a}\frac{\partial X^\nu}{\partial \xi^b}[g_{\mu\nu}(X) + \alpha' B_{\mu\nu}(X)]\right)}. \tag{19.17}$$

In this way we have shown the coupling of the D-brane with all the massless modes of the closed bosonic string. But we have only shown how it couples to ϕ^i, the massless scalars on the worldvolume, whereas there are other massless open string fields. In particular, as we saw in general for Neumann boundary conditions in the last chapter, there are massless vectors coming from $\alpha^i_{-1}|0,\vec{k}\rangle$, and on the D-brane we have the same boundary conditions. The mode A_a should couple gauge invariantly to the D-brane, i.e., through F_{ab}, and for small fields, we expect to have the usual Maxwell action $-\frac{1}{4}F_{ab}F^{ab}$. Moreover, it should fit inside the determinant.

Luckily, there is such a simple action, the Born–Infeld action for electromagnetism found in the 1930s, as an action without the singularities of the Maxwell action, that has a maximum electric field (and never reaches infinite values). It amounts to adding $\alpha' F_{\mu\nu}$ to the metric inside the determinant, so the full Dirac–Born–Infeld (DBI) action is

$$S_{DBI,Dp} = -T_p \int d^{p+1}\xi e^{-\phi} \sqrt{-\det\left[\frac{\partial X^\mu}{\partial \xi^a}\frac{\partial X^\nu}{\partial \xi^b}(g_{\mu\nu}(X) + \alpha' B_{\mu\nu}(X)) + \alpha' F_{ab}\right]}. \tag{19.18}$$

The action at $B_{ab} = X^i = \phi = 0$ is the Born–Infeld action, and the action for the scalars X^i's written before was defined (and studied) also by Dirac.

To the DBI action one in general needs to add another term, the "Wess–Zumino" (WZ) term that describes the coupling to antisymmetric tensor fields, called Ramond–Ramond (RR) fields, that appear for the supersymmetric string. There, the product of two spinorial (Ramond, or R) modes is decomposed in spacetime antisymmetric tensors $A_{\mu_1\ldots\mu_{p+1}}$.

The coupling to them is of the type current times gauge field, like in electromagnetism. There, we have the $\int d^d x j^\mu(\vec{x})A_\mu(\vec{x})$ term, for an electron source $j^0(\vec{x}) = e\delta^{(3)}(\vec{x} - \vec{x}(\tau))$, giving $\int d\tau A_0(\vec{x}(\tau))$. Now similarly, the "gauge field" $A_{\mu_1\ldots\mu_{p+1}}(x)$ is sourced by a $p+1$ dimensional brane, which gives a source

$$j^{01\ldots p}(\vec{x}) = \mu_p \delta^{(D-p-1)}(\vec{x} - \vec{x}(\xi)) \ , \tag{19.19}$$

so the simplest WZ term is

$$\int d^D x j^{\mu_1\ldots\mu_{p+1}}(\vec{x})A_{\mu_1\ldots\mu_{p+1}}(\vec{x}) = \mu_p \int d^{p+1}\xi A_{01\ldots p}(\vec{x}(\xi)). \tag{19.20}$$

Here μ_p is charge density (charge per unit p-dimensional volume) of the D-brane (in fact $\mu_p = T_p e^{\phi-\phi_0}$). This term is the leading part of the WZ term. In the supersymmetric theory, on a supersymmetric background, this term gives a $+T_p$ term that cancels the $-T_p$ constant in the expansion of the DBI term. In fact, through the presence of the $-T_p$ term in the DBI term we can see the need for another term in the action for the supersymmetric case, since supersymmetry requires zero energy for the ground state.

19.2 Chan–Paton Factors and Several D-Branes

We have seen how to write the full action for a D-brane, but we can in fact have more than one D-brane. We introduce them through so-called Chan–Paton factors. Indeed, it was noticed that we can add labels $|i\rangle$ to each of the open string endpoints, having thus $|i\rangle|j\rangle$, with $i, j = 1, \ldots, N$. But then one endpoint can be thought of being as in the $\mathbf{N}$ representation of $U(N)$, and the other in the $\bar{\mathbf{N}}$ representation, which means that the string itself can be thought to be in the $\mathbf{N} \times \bar{\mathbf{N}} = \text{adj}$, the adjoint representation of $U(N)$. Using the basis of adjoint matrices λ^a_{ij}, the wavefunctions of the states are

$$|\vec{k}, a\rangle = \sum_{i,j=1}^{N} |\vec{k}, ij\rangle \lambda^a_{ij}. \tag{19.21}$$

For the massless vector states, the fact that the Chan–Paton factors give a state in the adjoint representation means that we have nonabelian gauge fields.

In the case of Neumann and Dirichlet boundary conditions, giving D-branes, the index labels D-branes, so there are N D-branes, on which lives a $U(N)$ gauge theory. At the quadratic level, we have the action (already assuming the cancellation of the constant term through a supersymmetric WZ term)

$$S = \int d^{p+1}\xi(-2T_p)\,\mathrm{Tr}\left[-\frac{1}{2}D_a\vec{\phi}\cdot D^a\vec{\phi} - \frac{1}{4}F_{ab}F^{ab}\right]. \tag{19.22}$$

19.3 Quantizing Open Strings on Dp-Branes

We consider open string stretching between 2 parallel Dp-branes, situated at transverse positions, x^i_1 and x^i_2, i.e., with Dirichlet boundary conditions for the DD directions

$$X^i(\tau, \sigma = 0) = x^i_i; \quad X^i(\tau, \sigma = \pi) = x^i_2. \tag{19.23}$$

We first review what we did in the usual case, for NN directions. The Virasoro constraints were solved in light-cone coordinates by

$$\dot{X}^- \pm X'^- = \frac{1}{2\alpha'}\frac{1}{2p^+}(\dot{X}^I \pm X'^I)^2 = \frac{1}{2\alpha'}\frac{1}{2p^+}[(\dot{X}^i \pm X'^i)^2 + (\dot{X}^a \pm X'^a)^2]. \tag{19.24}$$

Here we have split the lightcone transverse coordinates as $X^I = (X^a, X^i)$ as before. But from solving the wave equation for the NN coordinates, expanding the solution in modes, we found

$$\dot{X}^a \pm X'^a = \sqrt{2\alpha'}\sum_{n\in\mathbb{Z}} \alpha_n^a e^{-in(\tau\pm\sigma)}. \tag{19.25}$$

Then substituting in the Virasoro constraint above, and expanding in modes, we found in particular the mass-shell condition from the $n = 0$ mode of the equation. The quantization of the modes was

$$[\alpha_n^a, \alpha_m^b] = m\eta^{ab}\delta_{m+n,0}. \tag{19.26}$$

We now turn to the DD coordinates and try to find a similar formula. Again, the wave equation for the X^i coordinates is

$$\partial_+\partial_- X^i = 0. \tag{19.27}$$

As before, the most general solution is a sum of a general function of $\tau + \sigma$ and a general function of $\tau - \sigma$,

$$X^i = f^i(\tau + \sigma) + g^i(\tau - \sigma). \tag{19.28}$$

Imposing the boundary condition at $\sigma = 0$, we find that $g^i(x) = x_1^i - f^i(x)$, so

$$X^i(\tau, \sigma) = x_1^i + f^i(\tau - \sigma) - f^i(\tau - \sigma). \tag{19.29}$$

Imposing also the boundary condition at $\sigma = \pi$, we find

$$f^i(\tau + \pi) - f^i(\tau - \pi) = x_2^i - x_1^i \Rightarrow f^i(x + 2\pi) = f^i(x) + x_2^i - x_1^i. \tag{19.30}$$

This periodicity condition means that we can expand the functions in sines and cosines, plus a linear term,

$$f^i(x) = \frac{x_2^i - x_1^i}{\pi} + \sum_{n\geq 1}(f_n^i(x)\sin n\sigma + \tilde{f}_n^i \cos n\sigma). \tag{19.31}$$

In turn, we find then the coordinates

$$X^i(\tau, \sigma) = x_1^i + \frac{x_2^i - x_1^i}{\pi}\sigma + \sum_{n\geq 1}(f_n^i \cos n\tau - \tilde{f}_n^i \sin n\tau)\sin n\sigma. \tag{19.32}$$

Redefining the coefficients in order to find $e^{+in\tau}$ and $e^{-in\tau}$ modes, we find

$$X^i(\tau,\sigma) = x_1^i + \frac{x_2^i - x_1^i}{\pi}\sigma + \sqrt{2\alpha'}\sum_{n\geq 0}\frac{1}{n}\alpha_n^i e^{-in\tau}\sin n\sigma. \tag{19.33}$$

We can include the $n = 0$ modes in the sum by defining

$$\sqrt{2\alpha'}\alpha_0^i = \frac{x_2^i - x_1^i}{\pi}. \tag{19.34}$$

Note that this is different than the NN case, in which case α_0^a were related to the momenta p^a. In the DD case, since the endpoints are fixed, there is no center of mass momentum for the string, so α_0^i is written instead in terms of the D-brane positions. With this definition, we find

$$X'^i \pm \dot{X}^i = \sqrt{2\alpha'}\sum_{n\in\mathbb{Z}}\alpha_n^i e^{-in(\tau\pm\sigma)}\ , \tag{19.35}$$

which up to an irrelevant overall sign is the same as we found in the NN case. That means that replacing in the Virasoro constraint we find the same relation as a function of α_n^i.

Then we can quantize in the same way as for the NN directions, with

$$[\alpha_n^i, \alpha_m^j] = m\delta^{ij}\delta_{m+n,0}. \tag{19.36}$$

Finally, that means that the mass-shell condition, coming from the modes of the Virasoro constraint, is the same when written in terms of α_n^i and α_n^a, the only difference being that α_0^i is now not related to $p^i = 0$. We thus obtain

$$2p^+p^- = \frac{1}{\alpha'}\left[\alpha' p^a p^a + \frac{1}{2}\alpha_0^i\alpha_0^i + \sum_{n\geq 1}(\alpha_{-n}^a\alpha_n^a + \alpha_{-n}^i\alpha_n^i) - 1\right]. \tag{19.37}$$

This can be rewritten as an equation for M^2,

$$\begin{aligned} M^2 \equiv 2p^+p^- - p^a p^a &= \frac{1}{2\alpha'}\alpha_0^i\alpha_0^i + \frac{1}{\alpha'}\left[\sum_{n\geq 1}(\alpha_{-n}^a\alpha_n^a + \alpha_{-n}^i\alpha_n^i) - 1\right] \\ &= \left(\frac{x_2^i - x_1^i}{2\pi\alpha'}\right)^2 + \frac{1}{\alpha'}(N^\perp - 1)\ , \end{aligned} \tag{19.38}$$

where the transverse number operator is

$$N^\perp \equiv \sum_{n\geq 1}\sum_a n a_n^{a\dagger}a_n^a + \sum_{m\geq 1}\sum_i m a_m^{i\dagger}a_m^i. \tag{19.39}$$

The extra term in M^2 is understood as coming from the string extended between the 2 D-branes. Since the energy of the string is $E = TL$, and $T = 1/(2\pi\alpha')$ is the string tension and L is the length of the string, we obtain exactly the extra term.

The ground state of the string is obtained for $N^\perp$ having eigenvalue 0, with momentum and Chan–Paton factors $[\tilde{i}\,\tilde{j}]$, $\tilde{i}$, $\tilde{j} = 1, 2$ for the 2 D-branes,

$$|p^+, p^a; [\tilde{i}\,\tilde{j}]\rangle \ , \tag{19.40}$$

and it has the mass squared of

$$M^2 = -\frac{1}{\alpha'} + \left(\frac{x_2^1 - x_1^i}{2\pi\alpha'}\right)^2 . \tag{19.41}$$

The general state is obtained by acting with creation operators onto it, so has the form

$$|\psi\rangle = \left[\prod_n \prod_i \frac{(a_n^{a\dagger})^{N_{an}}}{\sqrt{N_{a_n}!}}\right]\left[\prod_m \prod_a \frac{(a_m^{i\dagger})^{N_{im}}}{\sqrt{N_{i_m}!}}\right]|p^+, p^a; [\tilde{i}\,\tilde{j}]\rangle. \tag{19.42}$$

We see that the scalar ground state is not necessarily tachyonic anymore, it can be massless for

$$|x_2^i - x_1^i| = 2\pi\sqrt{\alpha'}. \tag{19.43}$$

On the other hand, the next excited states are now

$$\phi^{i[12]} \equiv a_1^{i\dagger}|p^+, p^a; [12]\rangle \ , \quad A_a^{[12]} \equiv a_1^{a\dagger}|p^+, p^a; [12]\rangle \ , \tag{19.44}$$

and correspond to the gauge field A_a and the transverse scalars (positions of the D-brane) ϕ^i, with Chan–Paton labels 1, 2 and with

$$M^2 = \left(\frac{x_2^1 - x_1^i}{2\pi\alpha'}\right)^2 . \tag{19.45}$$

If we consider coincident branes, these excited states are massless, and the vectors

$$a_1^{a\dagger}|p^+, p^a; [\tilde{i}\,\tilde{j}]\rangle \tag{19.46}$$

are Yang–Mills fields for a $U(N)$ group, while the scalars are charged adjoint scalars

$$a_1^{i\dagger}|p^+, p^a; [\tilde{i}\,\tilde{j}]\rangle. \tag{19.47}$$

Note that the scalars being in the adjoint, there are N^2 of them. But if we separate the branes, only the diagonal ones remain massless (corresponding to the unbroken $U(1)^N$ gauge group on the N D-branes), and off-diagonal ones get a mass, "eating"

the corresponding scalars via the Higgs mechanism. In this way, only the N diagonal scalars, corresponding to the N D-branes positions, remain massless, while the other ones become massive via the Higgs mechanism.

But when the branes are coincident, we have N^2 "coordinates" for the N D-branes, which means that classical geometry loses its meaning, and we have some sort of quantum geometry, where the space becomes matriceal.

Important Concepts to Remember

- D-branes are endpoints of open strings with Dirichlet boundary conditions in $D-p-1$ directions and Neumann boundary conditions in $p+1$ directions.
- They are dynamical objects in string theory, with well defined tension and charge. Their worldvolume modes can be excited by a mode coming in from the bulk of spacetime.
- Their worldvolume action is the sum of a DBI (Dirac–Born–Infeld) and a WZ (Wess–Zumino) or CS term.
- The DBI term is $S_{DBI} = -T_p \int e^{-\phi} \sqrt{-\det(g_{ab} + \alpha' B_{ab} + \alpha' F_{ab})}$, where g_{ab} and B_{ab} are the pullback onto the worldvolume of the spacetime fields, and F_{ab} is the field strength of the YM fields on the worldvolume.
- In a static gauge, the transverse coordinates become worldvolume scalars ϕ^i, and one can find the couplings of the worldvolume fields A_a, ϕ^i with the spacetime fields $g_{\mu\nu}$, $B_{\mu\nu}$, ϕ, etc.
- The WZ term has the coupling of the spacetime RR fields with the worldvolume of the D-brane. At leading order, it is $\mu_p \int A_{01\ldots p}$.
- When we have several D-branes on top of each other, this corresponds to Chan–Paton factors on the strings, and Yang–Mills massless fields.
- The mass of the open strings stretched between two D-branes has a term involving the distance between the D-branes, $M^2 = \frac{(\Delta x)^2}{2\pi\alpha'} + \frac{1}{\alpha'}(N^\perp - 1)$.

Further reading: See [20, 21] for more details.

Exercises

(1) Consider a Dp-brane moving in a general space (of a doubly-Wick rotated nonextremal Dp-brane), with fields

$$\begin{aligned} ds^2 &= H_p^{-1/2}(r)(-dt^2 + d\vec{x}_{p-1}^2 + f(r)dx_p^2) + H_p^{1/2}\left(\frac{dr^2}{f(r)} + r^2 d\Omega_{8-p}^2\right) \\ e^{2\phi} &= g_s^2 H_p^{-\frac{p-3}{2}} \\ C_p &= \frac{1}{g_s} H_p^{-1} dt \wedge dx^1 \wedge \cdots \wedge dx^p. \end{aligned} \tag{19.48}$$

Write a general formula for the potential $V_p(r)$ felt by the Dp-brane.
(2) Find the coupling between 2 gauge fields A_a, A_b and a dilaton ϕ in the D-brane action.

(3) Find the first 2 excited states for a string stretching between 2 D-branes situated at a distance $L = 4\pi\sqrt{\alpha'}$.
(4) Assuming that the nonabelian version of the DBI action, for X^I (transverse coordinates) and A_a that are $N \times N$ matrices, is obtained as a symmetrized trace of the usual DBI action, but for the nonabelian fields, write down explicitly this action up to quartic level in the fields (expanding the square root of the determinant). [Note: up to quartic order in the fields, this action is the correct D-brane action, but not up to arbitrary order.]

Chapter 20
Electromagnetic Fields on D-Branes and $\mathcal{N} = 4$ SYM. T-Duality of Closed Strings

In this chapter we continue the description of the field theory on D-branes, and we describe an important symmetry of string theory called T-duality, in the closed string case.

We have seen that the theory on the D-branes involves gauge fields A_μ^a of $U(N)$ and adjoint scalars ϕ^{ia}, with the standard quadratic action

$$S_{\text{lin.}} = -\frac{1}{g_{p+1}^2}\int d^{p+1}x(-2)\,\text{Tr}\left[\frac{1}{2}D_\mu\phi^i D^\mu\phi^i + \frac{1}{4}F_{\mu\nu}F^{\mu\nu}\right] \tag{20.1}$$

Here we have defined the trace with the normalization $\text{Tr}[T_a T_b] = -(1/2)\delta_{ab}$, and $D_\mu = \partial_\mu + g[A_\mu, .]$.

20.1 Supersymmetric String Theory

Until now, we have considered only the bosonic string, but as we have seen, the bosonic string has tachyons, so at least perturbatively it is unstable. Because of this, it is useful to consider a supersymmetric string theory.

There are several ways to obtain it. We are interested in a theory that has supersymmetry in spacetime, but it is easier to work with a theory with supersymmetry on the worldsheet. This is the Neveu–Schwarz–Ramond (NSR) formulation, or "spinning string". The spectrum turns out to be a truncation ("GSO projection") of the bosonic case, plus some superpartners; this truncation removes the tachyon, but keeps the massless states, and obtains a spacetime supersymmetric spectrum. There is also a formulation with spacetime supersymmetry, called the Green–Schwarz formulation, or the superstring; after gauge fixing a certain symmetry ("kappa symmetry"), it has also a worldsheet supersymmetry. It turns out to have the same spectrum as the NSR formulation, so is actually the same theory. There is also a formulation involving "pure spinors" (spinors involving some constraints on them), for the Berkovits

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_20

superstring, introduced in order to quantize covariantly the superstring (string with spacetime supersymmetry).

In all these formulations, one finds, in an analogous way to what we found in the bosonic case, that we need to have $D = 10$ spacetime dimensions in order to have a sensible quantum theory.

We have seen that the simplest example of *on-shell* supersymmetry (there is also *off-shell* supersymmetry that one can define by introducing some auxiliary fields) lives in 1+1 dimensions, of one real scalar ϕ and one Majorana fermion ψ, also with one on-shell degree of freedom. The *free* action was

$$S = -\frac{1}{2}\int d^2x[(\partial_\mu\phi)^2 + \bar{\psi}\partial\!\!\!/\psi] \,, \tag{20.2}$$

and it had supersymmetry

$$\begin{aligned} \delta\phi &= \bar{\epsilon}\psi \\ \delta\psi &= \partial\!\!\!/\phi\epsilon. \end{aligned} \tag{20.3}$$

Here the dimensions were $[\phi] = 0$, $[\psi] = 1/2$, $[\epsilon] = -1/2$. The NSR string has 10 scalars X^μ in 1+1 dimensions; the physical modes are the transverse X^I, $I = 1, \ldots, 8$ only, so one needs 8 on-shell fermionic degrees of freedom to have on-shell supersymmetry, but the action is constructed by taking multiples of the above one, reorganizing the fermions, and adding interactions.

20.2 D3-Brane Action and $\mathcal{N} = 4$ SYM

In superstring theory, the D-branes are supersymmetric (this could be proven by analyzing the boundary conditions on the open superstring, giving the D-brane theory). We will focus on D3-branes, since their worldvolume is 3+1 dimensional, so they are most relevant for us. The D3-branes, like the whole superstring theory, will live in 10 dimensions.

But as we saw, the gauge theory on them has the gauge fields A^a_μ, with 2 on-shell degrees of freedom, and the 6 scalars ϕ^i, $i = 1, \ldots, 6$, corresponding to the 6 coordinates transverse to the D3-brane, for a total of 8 on-shell bosonic degrees of freedom. Since in 4 dimensions, a minimal spinor has 2 on-shell degrees of freedom (the number of components of a Dirac fermion is $2^{[D/2]}$, which on-shell gives for the *minimal spinor* one real component in 2 dimensions, and two real components in 4 dimensions), we need to introduce 4 spinors, ψ^I, $I = 1, \ldots, 4$. The spinors will transform under $SU(4)$ acting on the index I. Since $SU(4) = SO(6)$, we can write the fundamental representation of $SO(6)$, to which ϕ^i belongs, as the antisymmetric representation $[IJ]$ of $SU(4)$, thus replacing $\phi^i \to \phi^{[IJ]}$. Since a supersymmetry transforms the gauge field into a spinor, we can guess that the theory will have $\mathcal{N} = 4$ supersymmetries, each (with a parameter ϵ^I) transforming A^a_μ into a different ψ^{aI}.

It turns out that for theories with spins ≤ 1 in 4 dimensions, there is a maximum number of supersymmetries equal to 4. The reason is that for $s \leq 1$, we have maximum helicity $+1$ and minimum -1, and one supersymmetry changes the helicity by 1/2. The same argument for spins $s \leq 2$, i.e. for supergravity theories, with helicities $\in [-2, +2]$, gives a maximum number of 8 supersymmetries. The gauge theory with $\mathcal{N} = 4$ supersymmetries, and with a maximum of two derivatives (usual kinetic terms), is actually unique, so the "linearized" (or more precisely, quadratic in derivatives) action on the D3-branes is this $\mathcal{N} = 4$ **SYM theory**.

The supersymmetry transformation rules for the bosons are easy to write down, based on the basic transformation $\delta\phi = \bar{\epsilon}\psi$, generalized first to $\delta\psi = \bar{\epsilon}_I \psi^I$ for 4 susies, and then to our case, when we just need to match indices (and as a result, to introduce a gamma function, i.e., a constant matrix with a μ index, for the gauge field), as

$$\begin{aligned} \delta A^a_\mu &= \bar{\epsilon}_I \gamma_\mu \psi^{Ia} \\ \delta\phi^{[IJ]} &= \frac{i}{2}\epsilon^{[I}\psi^{J]}. \end{aligned} \tag{20.4}$$

The fermion transformation law is more complicated, and will not be written here. The action for $\mathcal{N} = 4$ SYM is then

$$\begin{aligned} S = \int d^4x(-2)\,\mathrm{Tr}\Big[&-\frac{1}{2}D_\mu\phi_{[IJ]}D^\mu\phi^{[IJ]} - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} - \frac{1}{2}\bar{\psi}_I \not{D}\psi^J \\ &-g\bar{\psi}^I[\phi_{[IJ]}, \psi^J] - \frac{g^2}{4}[\phi_{[IJ]}, \phi_{[KL]}][\phi^{[IJ]}, \phi^{[KL]}]\Big]. \end{aligned} \tag{20.5}$$

The terms on the first line are just the kinetic terms (for the "quadratic" action), and the terms on the second line are necessary in order to maintain supersymmetry once we have introduced the minimal coupling of the gauge fields to the scalars and fermions. In fact, this action can be simply obtained from the action for $\mathcal{N} = 1$ SYM in 10 dimensions, which is just the minimally coupled action of a fermion coupled to a gauge fields,

$$S = \int d^{10}x(-2)\,\mathrm{Tr}\left[-\frac{1}{4}F_{MN}F^{MN} - \frac{1}{2}\bar{\psi}\not{D}\psi\right], \tag{20.6}$$

under dimensional reduction. Indeed, then the gauge field splits as $A_M = (A_\mu, \phi_i)$, and the field strength as $F_{MN} = (F_{\mu\nu}, F_{\mu i} = D_\mu\phi_i, F_{ij} = g^2[\phi_i, \phi_j])$, with $\phi_i \leftrightarrow \phi_{[IJ]}$, and the spinor splits as $\psi_A = \psi_{\alpha I}$.

We have only shown the action quadratic in derivative on the N D3-branes, namely $SU(N)$ $\mathcal{N} = 4$ SYM. But while for the $U(1)$ case the nonlinear DBI action is known, for the non-Abelian case, the full action is not known. For a while there was a conjecture that the result was the symmetric trace of the nonabelian version of the DBI action, but it turned out not to be the case beyond a few orders.

20.3 Nonlinear Born–Infeld Action

We therefore move on to some considerations about the nonlinear $U(1)$ action. In this case the full nonlinear supersymmetric action is known (it was found by M. Aganagic, C. Popescu and J. Schwarz), though it is quite complicated, and will not be reproduced here.

We consider the Born–Infeld action, the action originally found by Born and Infeld, in terms of only the Maxwell field (electromagnetism), i.e., the DBI action for $e^{\phi} = e^{\langle\phi\rangle} = g_s$, $\phi^i = 0$, $B_{\mu\nu} = 0$,

$$S_{BI} = -T_3 \int d^4x \left[\sqrt{-\det\left(\eta_{ab} + 2\pi\alpha' F_{ab}\right)} - 1\right] , \tag{20.7}$$

where $1/g_s$ was absorbed in T_3 and we have considered already the -1 coming from the WZ term. Then we write explicitly in terms of $E_i = F_{0i}$ and $B_i = \frac{1}{2}\epsilon_{ijk}F_{jk}$ the matrix

$$M_{ab} = \eta_{ab} + 2\pi\alpha' F_{ab} = \begin{pmatrix} -1 & -E_1 & -E_2 & -E_3 \\ E_1 & 1 & B_3 & -B_2 \\ E_2 & -B_3 & 1 & B_1 \\ E_3 & B_2 & -B_1 & 1 \end{pmatrix} , \tag{20.8}$$

where in the last equality we have put $2\pi\alpha' = 1$ for simplicity. Taking the determinant, we find

$$-\det M_{ab} = 1 - (\vec{E}^2 - \vec{B}^2) - (\vec{E}\cdot\vec{B})^2 , \tag{20.9}$$

and on the other hand we have, defining the dual electromagnetic field $\tilde{F}^{\mu\nu} = \frac{1}{2}\epsilon^{\mu\nu\rho\sigma}F_{\rho\sigma}$,

$$\begin{aligned} \frac{1}{2}(\vec{E}^2 - \vec{B}^2) &= -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} \\ \vec{E}\cdot\vec{B} &= -\frac{1}{8}\epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma} \equiv -\frac{1}{4}\tilde{F}_{\mu\nu}F^{\mu\nu}. \end{aligned} \tag{20.10}$$

The D3-brane action then becomes (reinstating $2\pi\alpha'$)

$$S_{BI,D3} = -T_3 \int d^4x \left[\sqrt{1 + \frac{1}{2}(2\pi\alpha')^2 F_{\mu\nu}F^{\mu\nu} - (2\pi\alpha')^4\left(\frac{1}{4}\tilde{F}_{\mu\nu}F^{\mu\nu}\right)^2} - 1\right] . \tag{20.11}$$

We note that by expanding to first nontrivial order, we get an overall factor of $-\frac{T_3(2\pi\alpha')^2}{4}$, to be identified with $-\frac{1}{4g_{3+1}^2}$ as in the linearized action (20.1). Therefore we obtain the D3-brane tension in terms of the SYM coupling as

$$T_3 = \frac{1}{(2\pi\alpha')^2 g_{3+1}^2}. \tag{20.12}$$

The Born–Infeld action for the D3-brane, in terms of electric and magnetic fields is then

$$S_{BI} = -\frac{1}{(2\pi\alpha')^2 g_{3+1}^2} \int d^4x \left[\sqrt{1 - (2\pi\alpha')^2(\vec{E}^2 - \vec{B}^2) - (2\pi\alpha')^4(\vec{E}\cdot\vec{B})^2} - 1 \right]. \tag{20.13}$$

At zero magnetic field, the purely electric action is

$$S_{BI} = -\frac{1}{(2\pi\alpha')^2 g_{3+1}^2} \int d^4x \left[\sqrt{1 - (2\pi\alpha')^2 \vec{E}^2} - 1 \right], \tag{20.14}$$

and we see that it admits a maximum electric field,

$$|\vec{E}| \leq \frac{1}{2\pi\alpha'} \equiv E_{\max}. \tag{20.15}$$

This was part of the motivation of Born and Infeld, who wanted a nonlinear action for electromagnetism that reduces to the usual Maxwell one for small field, but doesn't have singularities. An electron solution in Maxwell electromagnetism has infinite electric field at the electron source, but in BI electromagnetism we have only $E_{\max}$.

20.4 Closed Strings on Compact Spaces

With the purpose of manifesting an important symmetry of string theory, that will be of use in cosmology, called T-duality, we examine closed strings on compact spaces.

The Circle S^1

The simplest compact space is a circle, so consider that X^{25} is compactified on a circle, i.e., that we have the identification

$$X^{25} \sim X^{25} + 2\pi R. \tag{20.16}$$

On the other hand, closed strings are periodic in σ, so they obey

$$X^\mu(\tau, \sigma + 2\pi) = X^\mu(\tau, \sigma). \tag{20.17}$$

But because of the nontrivial identification, in the direction X^{25} we can have periodicity only up to an identification, i.e.,

$$X^{25}(\tau, \sigma + 2\pi) = X^{25}(\tau, \sigma) + 2\pi m R. \tag{20.18}$$

The interpretation of this is that the closed string winds m times around the compact circle direction. We will define the *winding*

$$w = \frac{mR}{\alpha'}. \tag{20.19}$$

This winding means that the mode expansion of X^{25} will get modified. As before, X^{25} satisfies a wave equation

$$\partial_+ \partial_- X^{25} = 0 \ , \tag{20.20}$$

solved by a sum of an arbitrary function of left movers and an arbitrary function of right movers,

$$X^{25}(\tau, \sigma) = X_L^{25}(\tau + \sigma) + X_R^{25}(\tau - \sigma). \tag{20.21}$$

We define the variables $u \equiv \tau + \sigma$ and $v = \tau - \sigma$. Then the winding-modified periodicity condition on X^{25} becomes

$$\begin{aligned} &X_L^{25}(u + 2\pi) + X_R^{25}(v - 2\pi) = X_L^{25}(u) + X_R^{25}(v) + 2\pi\alpha' w \Rightarrow \\ &X_L^{25}(u + 2\pi) - X_L^{25}(u) = X_R^{25}(v) - X_R^{25}(v - 2\pi) + 2\pi\alpha' w. \end{aligned} \tag{20.22}$$

But the left hand side is a function of u, and the right hand side is a function of v, so the solution is periodicity for both $X_L^{25}(u)$ and $X_R^{25}(v)$, which therefore can be expanded as usual in exponentials, but only up to linear term, i.e.,

$$\begin{aligned} X_L^{25}(u) &= x_{0,L}^{25} + \sqrt{\frac{\alpha'}{2}}\tilde{\alpha}_0^{25} u + i\sqrt{\frac{\alpha'}{2}} \sum_{n \neq 0} \frac{\tilde{\alpha}_n^{25}}{n} e^{-inu} \\ X_R^{25}(v) &= x_{0,R}^{25} + \sqrt{\frac{\alpha'}{2}}\alpha_0^{25} v + i\sqrt{\frac{\alpha'}{2}} \sum_{n \neq 0} \frac{\alpha_n^{25}}{n} e^{-inv} \ , \end{aligned} \tag{20.23}$$

with the constraint

$$\tilde{\alpha}_0^{25} - \alpha_0^{25} = \sqrt{2\alpha'} w. \tag{20.24}$$

We have seen that the momentum density was $\mathcal{P} = \dot{X}/(2\pi\alpha')$, so the momentum in the 25 direction is

$$p \equiv p^{25} = \frac{1}{2\pi\alpha'} \int_0^{2\pi} d\sigma (\dot{X}_L^{25} + \dot{X}_R^{25}). \tag{20.25}$$

Since $\dot{u} = \dot{v} = 1$, and the nonzero modes integrate to zero, from the explicit expressions for X_L^{25} and X_R^{25}, we find

$$p = \frac{\alpha_0^{25} + \tilde{\alpha}_0^{25}}{\sqrt{2\alpha'}}. \tag{20.26}$$

We invert these equations for w and p, to find

$$\alpha_0^{25} = \sqrt{\frac{\alpha'}{2}}(p - w); \quad \tilde{\alpha}_0^{25} = \sqrt{\frac{\alpha'}{2}}(p + w). \tag{20.27}$$

As we saw, $w = mR/\alpha'$ is written in terms of an integer m, but so is p. Indeed, on a compact space, e^{ipx} must be single valued around the circle, so $e^{2\pi i R p} = 1$, or

$$p = \frac{n}{R}. \tag{20.28}$$

Finally, substituting α_0^{25}, $\tilde{\alpha}_0^{25}$ in X^{25}, we find

$$X^{25}(\tau,\sigma) = X_L^{25}(\tau+\sigma) + X_R^{25}(\tau-\sigma) = x_0^{25} + \alpha' p\tau + \alpha' w\sigma + \sum_{n\neq 0} \frac{e^{-in\tau}}{n}(\alpha_n^{25} e^{in\sigma} + \tilde{\alpha}_n^{25} e^{-in\sigma}) , \tag{20.29}$$

where

$$x_0^{25} = x_{0,L}^{25} + x_{0,R}^{25}. \tag{20.30}$$

Virasoro Constraints and Mass Spectrum on Circle

Now we are in a position to calculate the mass spectrum of closed strings on the circle, from the Virasoro constraints. We remember that the modes of the Virasoro constraints are L_n and $\tilde{L}_n$, and in the $n = 0$ case we have $L_0 - 1 = \tilde{L}_0 - 1 = 0$, and from these latter ones we obtain the mass spectrum. The only difference with respect to the calculation in previous chapters is in the form of α_0^{25} and $\tilde{\alpha}_0^{25}$, but otherwise we have the same result. In lightcone coordinates $(I, 25) = (+, -, i, 25)$, we have as before

$$\begin{aligned} L_0 &= \frac{1}{2}(\alpha_0^I \alpha_0^I + \alpha_0^{25}\alpha_0^{25}) + N^\perp = \frac{\alpha'}{4}(-2p^+p^- + p^i p^i) + \frac{\alpha'}{4}(p-w)^2 + N^\perp \\ \tilde{L}_0 &= \frac{1}{2}(\tilde{\alpha}_0^I \tilde{\alpha}_0^I + \tilde{\alpha}_0^{25}\tilde{\alpha}_0^{25}) + \tilde{N}^\perp = \frac{\alpha'}{4}(-2p^+p^- + p^i p^i) + \frac{\alpha'}{4}(p+w)^2 + \tilde{N}^\perp. \end{aligned} \tag{20.31}$$

Then, as usual, the Hamiltonian constraint $H = L_0 + \tilde{L}_0 - 2 = 0$ gives the mass spectrum, and the (worldsheet) momentum constraint $P = L_0 - \tilde{L}_0$ gives the level matching condition.

The worldsheet momentum constraint is now

$$L_0 - \tilde{L}_0 = -\alpha' pw + N^\perp - \tilde{N}^\perp - 2 = 0 \Rightarrow N^\perp - \tilde{N}^\perp = \alpha' pw = nm , \tag{20.32}$$

so the new level matching is only up to the integer nm.

The mass spectrum is given by the condition $L_0 + \tilde{L}_0 - 2 = 0$, or

$$\alpha'(-2p^+p^- + p^i p^i) + \frac{\alpha'}{2}(p^2 + w^2) + N^\perp + \tilde{N}^\perp - 2 = 0 , \tag{20.33}$$

which implies that the mass squared in the compactified space is

$$M^2_{\text{compactified}} \equiv 2p^+p^- - p^i p^i = p^2 + w^2 + \frac{2}{\alpha'}(N^\perp + \tilde{N}^\perp - 2). \tag{20.34}$$

The extra term in the spectrum compared to the uncompactified case is the $w^2 = (mR/\alpha')^2$ term, which comes from the length of the wound string. Indeed, we have seen in the previous chapter that an open string stretched between two D-branes had a mass contribution coming from the tension times the length of the string. A similar thing happens here, since the contribution to the mass is

$$\Delta M = T\Delta L = \frac{1}{2\pi\alpha'} 2\pi m R = \frac{mR}{\alpha'} = w. \tag{20.35}$$

20.5 T-Duality of Closed Strings

We can write the spectrum in terms of integers as

$$M^2_{\text{compactified}} = \left(\frac{n}{R}\right)^2 + \left(\frac{mR}{\alpha'}\right)^2 + \frac{2}{\alpha'}(N^\perp + \tilde{N}^\perp - 2). \tag{20.36}$$

If we do so, we notice a symmetry of the spectrum, called *T-duality*, which exchanges n with m, and R with

$$\tilde{R} = \frac{\alpha'}{R}\ , \tag{20.37}$$

so also p with w, i.e., momentum with winding. More precisely, we find that

$$M^2_{\text{compactified}}(R; n, m) = M^2_{\text{compactified}}(\tilde{R}; m, n). \tag{20.38}$$

This is a symmetry that exists only in string theory, as it exchanges the momentum modes with winding modes, which only exist for the string. It says that a certain *vacuum*, characterized by a value of the radius R, is equivalent with another vacuum, characterized by another radius $\tilde{R}$. The radius is assumed to be a *modulus* in this case, i.e., to have no potential (perturbatively, non-perturbatively there is a potential), so it costs no energy to change it. Then, effectively, theories with $R < \sqrt{\alpha'} \equiv l_s$ (string length) are equivalent with theories with $R > l_s$, since when the radius becomes small, while the momenta have increasing gaps between them, the windings become continuous, just like the momenta are in the decompactified limit.

Exchanging n with m and R with $\tilde{R}$, thus p with w has the effect to act on the zero mode oscillators as

$$\tilde{\alpha}_0^{25} \leftrightarrow \tilde{\alpha}_0^{25}; \quad \alpha_0^{25} \leftrightarrow -\alpha_0^{25}. \tag{20.39}$$

But then we note that in fact, the T-duality symmetry of the spectrum is more general, since we can also exchange the nonzero mode in the same way as the zero modes, i.e., by

$$\tilde{\alpha}_n^{25} \leftrightarrow \tilde{\alpha}_n^{25}; \quad \alpha_n^{25} \leftrightarrow -\alpha_n^{25}. \tag{20.40}$$

Moreover, we can also exchange the constant pieces as

$$x_{0,L}^{25} \leftrightarrow x_{0,L}^{25}; \quad x_{0,R}^{25} \leftrightarrow -x_{0,R}^{25} \ , \tag{20.41}$$

so that finally the whole X_L^{25} and X_R^{25} are changed under the symmetry as

$$X_L^{25}(\tau+\sigma) \leftrightarrow X_L^{25}(\tau+\sigma); \quad X_R^{25}(\tau-\sigma) \leftrightarrow -X_R^{25}(\tau-\sigma) \ , \tag{20.42}$$

so $X^{25}(\tau,\sigma)$ is exchanged with

$$\begin{aligned} X'^{25}(\tau,\sigma) &= X_L^{25}(\tau+\sigma) - X_R^{25}(\tau-\sigma) \\ &= q_0^{25} + \alpha' w\tau + \alpha' p\sigma + i\sqrt{\frac{\alpha'}{2}} \sum_{n\neq 0} \frac{e^{-in\tau}}{n} (\tilde{\alpha}_n^{25} e^{-in\sigma} - \alpha_n^{25} e^{in\sigma}) \ , \end{aligned} \tag{20.43}$$

where

$$q_0^{25} = x_{0,L}^{25} - x_{0,R}^{25}. \tag{20.44}$$

We have defined T-duality only as a symmetry of the mass spectrum, i.e., of the free theory, but in fact is a symmetry of the interactions as well, so that in fact is a symmetry of the full string theory.

As such, it can also be considered as a symmetry of the theory in background created by the condensation of the string modes, i.e., in supergravity backgrounds. But then the string theory in some background will be related to a string theory in another background, related by the *Buscher rules*. The original Buscher rules, written for the background of the massless modes of the closed bosonic string, $(g_{\mu\nu}, B_{\mu\nu}, \phi)$, are given by (here 0 refers to the T-duality direction, not to time, and i refers to the other directions)

$$\begin{aligned} \hat{g}_{00} &= \frac{1}{g_{00}} \\ \hat{g}_{0i} &= \frac{B_{0i}}{g_{00}} \\ \hat{g}_{ij} &= g_{ij} - \frac{g_{0i}g_{0j} - B_{0i}B_{0j}}{g_{00}} \\ \hat{B}_{0i} &= \frac{g_{0i}}{g_{00}} \\ \hat{B}_{ij} &= B_{ij} + \frac{g_{0i}B_{0j} - g_{0j}B_{0i}}{g_{00}} \end{aligned}$$

$$\hat{\phi} = \phi - \frac{1}{2}\log g_{00}. \tag{20.45}$$

Note that the first relation is the usual $\tilde{R} = \alpha'/R$ for $\alpha' = 1$, since we can use a 2π periodicity, but put R^2 in the metric as g_{00}, so $\hat{g}_{00} = 1/g_{00}$ corresponds to $\hat{R}^2 = 1/R^2$.

T-duality can be generalized also for a general compact space K_n. In particular, for a torus $T^n = (S^1)^n$, we have T-duality on each circle.

Important Concepts to Remember

- Superstring theory has several formulations: Green–Schwarz has supersymmetry in spacetime and κ symmetry on worldsheet, which when gauge-fixed leads to worldsheet supersymmetry. Neveu–Schwarz–Ramond has supersymmetry on the worldvolume ("spinning string"). And Berkovits's "pure spinor" formulation is a way to covariantly quantize the string with spacetime supersymmetry.
- Supersymmetry can be defined on-shell or off-shell, and relates bosons with fermions, and degree of freedom with degree of freedom.
- The superstring lives in 10 dimensions, though a modification of the argument from the previous chapter about 26 dimensions for the bosonic string.
- D3-branes have a 3+1 dimensional worldvolume, on which the maximum number of supersymmetries (actually realized) for spins ≤ 1 is 4. The unique $\mathcal{N} = 4$ SYM theory is the linearized theory on the worldvolume of the D3-branes.
- The $\mathcal{N} = 4$ SYM has 6 scalars transforming in $SO(6) = SU(4)$, a gauge field and 4 fermions in the fundamental of $SU(4)$.
- There is a maximum electric field on the D3-brane, $E_{\max} = 1/(2\pi/\alpha')$.
- Closed strings on compact spaces have not just momentum $p = n/R$ but also winding $w = mR/\alpha'$.
- The mass spectrum of closed strings on a circle has *T-duality*, $M^2_{\text{compactified}}(R; n, m) = M^2_{\text{compactified}}(\tilde{R}; m, n)$, where $\tilde{R} = \alpha'/R$.
- The T-duality relation exchanges momentum with winding, n with m, and changes the sign of X_R, while leaving X_R invariant.
- T-duality has an action on the massless modes of the string, which implies a relation on the supergravity background fields, the Buscher rules.

Further reading: See [20–22] on string theory, as well as the more specialized [23] on D-branes.

Exercises

(1) Is there a maximum *magnetic* field on the brane? Perform Lorentz transformations and show consistency with a maximum electric field.
(2) Consider the closed string spectrum for compactification on the torus T^6 of type II string theory. Write down the mass formula for the string, and specialize it to $R_i = \sqrt{\alpha'}$.

(3) Calculate the solution T-dual (on X^1) to

$$\begin{aligned} ds^2 &= dx_1^2 + dx_2^2 + dx_3^2 \\ e^{\phi} &= e^{\phi_0} \\ B_{12} &= mx_3 \end{aligned} \tag{20.46}$$

using the Buscher rules.
(4) Show explicitly that the action of $\mathcal{N} = 4$ SYM in 4 dimensions is the dimensional reduction of the action for $\mathcal{N} = 1$ SYM in 10 dimensions.

Chapter 21
T-Duality of Open Strings. M-Theory and the Duality Web

In the last chapter, we considered T-duality for closed strings. Now we start off by considering T-duality for open strings, and see that it will lead to D-branes. Then, we define the low energy supergravity limits of string theory, the strong coupling limit of string theory known as M-theory, and finally show that all the various incarnations of string theory are related into a single duality web, as facets of the same unique theory.

21.1 T-Duality of Open Strings

We consider open strings with free endpoints, i.e., with Neumann boundary conditions on both endpoints (NN strings) for the compact direction X^{25}. Then the mode expansion for X^{25} is the one we already found for NN strings,

$$X^{25}(\tau,\sigma) = x_0^{25} + \sqrt{2\alpha'}\alpha_0^{25} + i\sqrt{2\alpha'}\sum_{n\neq 0}\frac{\alpha_n^{25}}{n}e^{-in\tau}\cos n\sigma \ , \tag{21.1}$$

just that the zero mode α_0^{25}, still expressible in terms of the momentum in the 25th direction, is now quantized,

$$\alpha_0^{25} = \sqrt{2\alpha'}p = \sqrt{2\alpha'}\frac{n}{R}. \tag{21.2}$$

Note that the strings are open and free, so there is no winding mode.

As before, the above formula can be actually found by splitting X^{25} into left- and right-movers, and then seeing that for the boundary condition we need to have $\alpha_n^{25} = \tilde{\alpha}_n^{25}$. In any case, we can split X^{25} as

$$X^{25}(\tau,\sigma) = X_L^{25}(\tau+\sigma) + X_R^{25}(\tau-\sigma) \ , \tag{21.3}$$

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_21

where each of the modes is given by the free closed string expansion,

$$X_L^{25} = \frac{x_0^{25} + q_0^{25}}{2} + \sqrt{\frac{\alpha'}{2}}\alpha_0^{25}(\tau+\sigma) + i\sqrt{\frac{\alpha'}{2}}\sum_{n\neq 0}\frac{\alpha_n^{25}}{n}e^{-in\tau}e^{-in\sigma}$$
$$X_R^{25} = \frac{x_0^{25} - q_0^{25}}{2} + \sqrt{\frac{\alpha'}{2}}\alpha_0^{25}(\tau-\sigma) + i\sqrt{\frac{\alpha'}{2}}\sum_{n\neq 0}\frac{\alpha_n^{25}}{n}e^{-in\tau}e^{+in\sigma}\ , \quad (21.4)$$

where we have introduced a q_0^{25}, since the zero modes of X_L and X_R are in principle independent.

Now we can do the same T-duality transformation that we did in the case of closed strings, i.e., replace $X^{25}(\tau,\sigma)$ with

$$X'^{25}(\tau,\sigma) = X_L^{25}(\tau+\sigma) - X_R^{25}(\tau-\sigma) = q_0^{25} + \sqrt{2\alpha'}\alpha_0^{25}\sigma + \sqrt{2\alpha'}\sum_{n\neq 0}\frac{\alpha_n^{25}}{n}e^{-in\tau}\sin n\sigma. \quad (21.5)$$

But now we notice that this is the same as the expansion for an open string stretching between 2 D-branes, just with

$$\alpha_0^{25} = \sqrt{2\alpha'}\frac{n}{R} = \frac{1}{\sqrt{2\alpha'}}\frac{x_2^{25} - x_1^{25}}{\pi}. \quad (21.6)$$

This means that the distance between the possible D-branes is

$$X'^{25}(\tau,\sigma=\pi) - X'^{25}(\tau,\sigma=0) = 2\alpha'\frac{n}{R} = 2\pi\tilde{R}n\ , \quad (21.7)$$

where in the last equality we have used the radius of the T-dual space, $\tilde{R} = \alpha'/R$. And we can also say that

$$X'^{25}(\tau,\sigma=0) = q_0^{25}\ , \quad (21.8)$$

which can be therefore identified with the position of the D-brane on the T-dual circle. The difference of $2\pi\tilde{R}n$ means that the T-dual string winds n times around the T-dual circle, and then comes back to the same D-brane at q_0^{25}.

We see then that under T-duality, a D25-brane (free Neumann boundary conditions for the whole space) turned into a D24-brane.

Moreover, the boundary conditions are interchanged (Neumann with Dirichlet), since

$$\partial_\sigma X^{25}(\tau,\sigma) = \frac{dX_L^{25}(u=\tau+\sigma)}{du} - \frac{dX_R^{25}}{du}(u=\tau-\sigma) = \partial_\tau X'^{25}(\tau,\sigma)$$
$$\partial_\tau X^{25}(\tau,\sigma) = \frac{dX_L^{25}(u=\tau+\sigma)}{du} + \frac{dX_R^{25}}{du}(u=\tau-\sigma) = \partial_\sigma X'^{25}(\tau,\sigma)\ , \quad (21.9)$$

which means that Neumann conditions become Dirichlet, and Dirichlet become Neumann,

$$\begin{aligned}\partial_\sigma X^{25}(\tau,\sigma) = 0 &\to \partial_\tau X^{25}(\tau,\sigma) = 0\\ \partial_\tau X^{25}(\tau,\sigma) = 0 &\to \partial_\sigma X^{25}(\tau,\sigma) = 0.\end{aligned} \tag{21.10}$$

Then we can say more generally, that if we take a Dp-brane and make a T-duality along a parallel direction to the brane, we turn the Dp into D$(p-1)$-brane, and if we make a T-duality along a transverse direction to the brane, we turn it into a D$(p+1)$-brane. Either way, the dimensionality of the brane changes.

21.2 T-Duality with Chan–Paton Factors: Several D-Branes

We have seen what happens when we perform T-duality on a empty space with a single open string: we obtain a D-brane. But we want to ask what happens when we perform T-duality on an open string with Chan–Paton factors for a group $U(N)$.

We need to introduce the concept of Wilson line, or loop (for a closed one) associated with a path in spacetime. For a closed path C and an abelian ($U(1)$) group, it is

$$W \equiv e^{iq\oint_C dx A_x} \equiv e^{iw}. \tag{21.11}$$

Here we have concentrated on a single coordinate x, which will be moreover chosen to be compact, and the contour C to wrap around it. In general, we would have $\int_C dx^\mu A_\mu$ in the exponent.

The Wilson loop is gauge invariant. Indeed, consider gauge transformations

$$A_x \to A_x + \partial_x \lambda\,, \tag{21.12}$$

and a periodic gauge parameter, modulo $2\pi m/q$,

$$q\lambda(x + 2\pi R) = q\lambda(x) + 2\pi m. \tag{21.13}$$

Then the gauge transformation $U = e^{iq\lambda} = 1$, and the variation in the exponent of the Wilson loop is

$$\delta q \oint_C dx A_x = \oint dx \partial_x \lambda = \lambda(2\pi R) - \lambda(0) = 2\pi m\,, \tag{21.14}$$

which means that the Wilson loop is invariant,

$$W \to W \cdot e^{2\pi i m} = W. \tag{21.15}$$

Moreover, because of the periodicity,

$$w \equiv q \oint dx A_x \equiv \theta \in [0, 2\pi] \tag{21.16}$$

is an angle.

Given a value for the the Wilson loop, it can be realized in the simplest way by a constant A_x, namely

$$q A_x = \frac{\theta}{2\pi R}. \tag{21.17}$$

But when adding a gauge coupling qA_μ to a particle, the momentum is changed to the gauge invariant form,

$$p_\mu \to p_\mu - q A_\mu \,, \tag{21.18}$$

so in our case

$$p_{25} \to p_{25} - q A_{25} \Rightarrow \frac{n}{R} \to \frac{n}{R} - \frac{\theta}{2\pi R}. \tag{21.19}$$

Consider now an open string with $U(N)$ Chan–Paton factors, and a general Wilson loop. It can be realized by a constant $U(N)$ matrix A_{25}, which can therefore be diagonalized by gauge transformations, to a form

$$A_{25} = \frac{1}{2\pi R} \mathrm{diag}(\theta_1, \ldots, \theta_N). \tag{21.20}$$

By its presence, we break the $U(N)$ invariance to only $U(1)^N$. But then the Chan–Paton state $|ij\rangle$ has charge $+1$ under $U(1)_i$ and charge -1 under $U(1)_j$ (since we can think of one end other string as belonging to the fundamental representation, and the other to the antifundamental; when considering a $U(1)$ subgroup, one corresponds to $+1$ charge, the other -1 charge). That means that the gauge invariant form of the momentum is now

$$\frac{2\pi n + \theta_i - \theta_j}{2\pi R} = \frac{1}{2\pi\alpha'} \tilde{R}(2\pi n + \theta_i - \theta_j). \tag{21.21}$$

Substituting this into the usual open string spectrum, we now find

$$M^2 = \left(\frac{2\pi n - (\theta_j - \theta_i)}{2\pi R} \right)^2 + \frac{1}{\alpha'}(N^\perp - 1). \tag{21.22}$$

We see then that now $\tilde{R}\theta_i$ plays the role of position of an ith D-brane in X'^{25} (which has radius $\tilde{R}$), and the T-dual picture is one where we have N D24-branes, at the locations θ_i, as in Fig. 21.1. Moreover, we can now consider

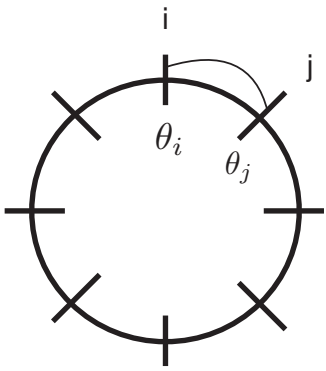

Fig. 21.1 The T-dual of Wilson lines is a set of D-branes at positions $R\theta_i$, $R\theta_j$ in the compact dimension, in between which we have open strings

$$X'^{25}(\tau,\sigma=\pi)-X'^{25}(\tau,\sigma=0)=\int_0^\pi d\sigma\partial_\sigma X'^{25}=\int_0^\pi d\sigma\partial_\tau X^{25}=2\pi\alpha' p$$
$$=2\pi\alpha'\left(\frac{n}{R}+\frac{\Delta\theta}{2\pi R}\right)=\tilde{R}(2\pi n+\Delta\theta)\ , \quad (21.23)$$

so indeed the string stretches between two D-branes at θ_i and θ_j, after winding n times around the circle.

21.3 Supergravity Actions

We want to study supersymmetric string theory, i.e., superstring theory, which lives in $D=10$ dimensions for consistency, as we saw. Moreover, it has $\mathcal{N}=2$ supersymmetry, which is the maximal possible in order to have spins $s\leq 2$ in 4 dimensions, since as we argued that implied $\mathcal{N}\leq 8$ in 4 dimensions, and one 10 dimensional minimal spinor has the same number of degrees of freedom as 8 spinors in 4 dimensions, i.e., 16. But if there are two supersymmetries, and each one relating the bosons to a different fermion, there are two fermions as well, which means that there are two types of theories, depending on the chirality (eigenvalue of $\Gamma_{D+1}=\Gamma_1\dots\Gamma_D$) of the spinors,

$$\Gamma_{11}\psi_i=\pm\psi_i. \quad (21.24)$$

- If the spinors have the same chirality,

$$\Gamma_{11}\psi_i=+\psi_i\ , i=1,2\ , \quad (21.25)$$

  we have the **type IIB** theory, and
- If the spinors have opposite chirality,

$$\Gamma_{11}\psi_i=(-1)^i\psi_i\ , i=1,2\ , \quad (21.26)$$

  we have **type IIA** theory.

At low energies, i.e., for $E \ll 1/\sqrt{\alpha'}$ (since the unique energy scale of string theory is made from α' in this way), the string theory turns into a supersymmetric theory that involves $g_{\mu\nu}$ (since as we saw the closed string massless modes contain $g_{\mu\nu}$, $B_{\mu\nu}$, ϕ), i.e., *supergravity*. Correspondingly, at low energy we have type IIA supergravity or type IIB supergravity.

In terms of the superstring modes, which as we said were composed of the products (for left- and right-movers) of either the bosonic Neveu–Schwarz (NS) sector or the fermionic Ramond (R) sector, we have the fields

- The NS-NS modes are bosons, and the massless modes are the same as the modes of the closed bosonic string, $(g_{\mu\nu}, B_{\mu\nu}, \phi)$. They are common to the IIA and IIB theories.
- The NS-R and the R-NS sectors are fermions, but we will not study their massless modes.
- The R-R modes are also bosons, but their massless modes differ between the IIA and IIB theories. Due to the fact that they are constructed, they are associated to objects of the type $\bar{\psi}\Gamma_{\mu_1\ldots\mu_{p+1}}\chi$, where ψ, χ are fermions, so they are antisymmetric tensor fields $A_{(p+1)}$.
  type IIA. In type IIA theory, we have odd form potentials: $A_{(1)}$ or A_μ, a usual gauge field, with field strength $F_{(2)}$ or $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$, which couples electrically to sources = D0-branes ("D-particles"). In general as we saw, Dp-branes have a WZ term that includes the electric-type coupling $\mu_p \int A_{(p+1)}$, so D-particles couple to gauge fields in the usual way. We also have a $A_{(3)}$, or $A_{\mu\nu\rho}$, with field strength $F_{\mu_1\ldots\mu_4} = 3\partial_{[\mu_1} A_{\mu_2\mu_3\mu_4]}$, and coupling electrically to D2-branes. We also have D4-branes, coupling magnetically (i.e., through the 5-form dual to $A_{(3)}$, $A_{(5)} = *A_{(3)}$) to $A_{(3)}$, and D6-branes, coupling magnetically to $A_{(1)}$.
  type IIB. In type IIB theory, we have even form potentials: the (pseudo)scalar (axion) a with field strength $F_{(1)}$ or $F_\mu = \partial_\mu a$, coupling electrically to a D(-1)-brane, or "D-instanton". This is an object with Dirichlet boundary conditions in all space *and time* directions. We also have a 2-form $A_{(2)}$ or $A_{\mu\nu\rho}$ with field strength $F_{(3)}$ or $F_{\mu\nu\rho}$ and coupling electrically to D1-branes ("D-strings"), and a self-dual 4-form $A^+_{(4)}$ or $A^+_{\mu\nu\rho\sigma}$, with field strength $F^+_{\mu_1\ldots\mu_5}$ ($F^+_{(5)}$) that is self dual,

$$F^+_{\mu_1\ldots\mu_5} = \frac{1}{5!}\epsilon_{\mu_1\ldots\mu_5}{}^{\mu_6\ldots\mu_{10}} F_{\mu_6\ldots\mu_{10}} \ , \tag{21.27}$$

  and that couples to a D3-brane. Note that we impose the self-duality condition in order to obtain an irreducible representation (if not, we can reduce the representation into self-dual and anti-self-dual parts). We also have D5-branes, coupling magnetically to $A_{(2)}$, and D7-branes, coupling magnetically to a. There is no magnetic D3-brane, since the D3-brane is self-dual (so both electric and magnetic).

We now write supergravity actions, in the so-called "string frame", which means that we consider instead of the "Einstein metric" $G^{(E)}_{\mu\nu}$, for which the gravity action is the Einstein–Hilbert term $\frac{1}{2\kappa_N^2}\int d^D x\sqrt{-G}R$, the action $\frac{1}{2\kappa_N^2}\int d^D x\sqrt{-G^{(S)}}e^{-2\phi}R$. This

is what one naturally obtains in string theory, since the $e^{-2\phi}$ factor has as a VEV $e^{-2\langle\phi\rangle} \equiv 1/g_s^2$, as expected from a (closed) string action.

In the string frame, the *bosonic type IIA supergravity action* is

$$\begin{aligned} S_{IIA}^{\text{bose}} = \frac{1}{2\kappa_{10}^2}\int d^{10}x \left\{\sqrt{-G}\left[e^{-2\Phi}\left(R + 4\partial_\mu\Phi\partial^\mu\Phi - \frac{1}{2}|H_3|^2\right)\right.\right. \\ \left. -\frac{1}{2}|F_2|^2 - \frac{1}{2}|\tilde{F}_4|^2\right] \\ \left. -\frac{1}{2}B_2 \wedge F_4 \wedge F_4\right\}, \end{aligned} \tag{21.28}$$

where

$$\begin{aligned} \tilde{F}_4 &\equiv dA_3 - A_1 \wedge F_3 \\ \int d^D x\sqrt{-G}|F_p|^2 &\equiv \int d^D x\sqrt{-G}\frac{1}{p!}G^{M_1N_1}\ldots G^{M_pN_p}F_{M_1\ldots M_p}F_{N_1\ldots N_p}, \end{aligned} \tag{21.29}$$

and the terms on the first line in the action, with $e^{-2\Phi}$, are the NS-NS terms (common to IIA and IIB), the terms on the second line are the R-R terms (just for IIA) and the term on the third line is a CS term.

The *bosonic type IIB supergravity action* in string frame has a caveat. Since the field $A_{(4)}^+$ is self-dual, there is no covariant 2-derivative action. One has either to write an action without the self-duality and make the self-duality condition a constraint to be added to the equations of motion, or one breaks explicit Lorentz covariance, or one introduces auxiliary fields. It is customary to present the first form, which is what we will do here. The action is then

$$\begin{aligned} S_{IIB}^{\text{bose}} = \frac{1}{2\kappa_{10}^2}\int d^{10}x \left\{\sqrt{-G}\left[e^{-2\Phi}\left(R + 4\partial_\mu\Phi\partial^\mu\Phi - \frac{1}{2}|H_3|^2\right)\right.\right. \\ \left. -\frac{1}{2}|F_1|^2 - \frac{1}{2}|\tilde{F}_3|^2 - \frac{1}{4}|\tilde{F}_5|^2\right] \\ \left. -\frac{1}{2}A_4 \wedge H_3 \wedge F_3\right\}, \end{aligned} \tag{21.30}$$

where

$$\begin{aligned} \tilde{F}_3 &\equiv dA_2 - A_0 \wedge H_3 \\ \tilde{F}_5 &\equiv F_5 - \frac{1}{2}A_2 \wedge H_3 + \frac{1}{2}B_2 \wedge F_3. \end{aligned} \tag{21.31}$$

The self-duality of $\tilde{F}_5$ is imposed as a constraint, in addition to the equations of motion,

$$\tilde{F}_5 = *\tilde{F}_5. \tag{21.32}$$

Note that the extra factor of 1/2 in front of the kinetic action for $\tilde{F}_5$ is due to the fact that in the action we have not imposed the self-duality constraint, so we have twice as many degrees of freedom as we should.

S-Duality of Type IIB Theory

We now describe an important duality symmetry of type IIB theory. We start with the IIB supergravity action, which is invariant under $Sl(2,\mathbb{R})$. To write the action in an invariant way, we first define the combination of the axion (pseudo)scalar C_0 and the scalar dilaton

$$\tau \equiv C_0 + ie^{-\Phi}\ , \tag{21.33}$$

and then out of it, the matrix

$$\mathcal{M}_{ij} \equiv \frac{1}{\mathrm{Im}\tau}\begin{pmatrix} |\tau|^2 & -\mathrm{Re}\tau \\ -\mathrm{Re}\tau & 1 \end{pmatrix}. \tag{21.34}$$

We also define the column vector of the NS-NS and R-R field strengths

$$F_3^i = \begin{pmatrix} H_3 \\ F_3 \end{pmatrix}. \tag{21.35}$$

We then write the action *in the Einstein frame*, i.e. in terms of the Einstein metric $G^{(E)}_{\mu\nu}$, as

$$\begin{aligned} S^{\mathrm{bose}}_{IIB} = \frac{1}{2\kappa_{10}^2}\int d^{10}x\sqrt{-G_E}\left[R_E - \frac{\partial_\mu\bar{\tau}\partial^\mu\tau}{2(\mathrm{Im}\tau)^2} - \frac{\mathcal{M}_{ij}}{2}F_3^iF_3^j - \frac{1}{4}|\tilde{F}_5|^2\right] \\ - \frac{\epsilon_{ij}}{8\kappa_{10}^2}\int d^{10}x A_4\wedge F_3^i\wedge F_3^j. \end{aligned} \tag{21.36}$$

In this way, the action is manifestly invariant under the $Sl(2,\mathbb{R})$ group transformation

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\ , \tag{21.37}$$

(with determinant one, $ad - bc = 1$) acting on τ by

$$\tau \to \tau' = \frac{a\tau + b}{c\tau + d}\ , \tag{21.38}$$

and on F_3^i by the rearranged matrix

$$\Lambda^i{}_j \equiv \begin{pmatrix} d & c \\ b & a \end{pmatrix}\ , \tag{21.39}$$

as

$$F_3^i = \Lambda^i{}_j F_3^j . \tag{21.40}$$

Under this transformation, the matrix $\mathcal{M}_{ij}$ transforms as

$$\mathcal{M} \to \mathcal{M}' = (\Lambda^{-1})^T \mathcal{M} \Lambda^{-1} , \tag{21.41}$$

making the F_3^i kinetic term invariant. Also F_5 and $G_{\mu\nu}^{(E)}$ are invariant.

In this way, the type IIB supergravity action is invariant, or rather the equations of motion are, since we need to impose the self-duality constraint on them. But in the full quantum string theory, only a $Sl(2,\mathbb{Z})$ subgroup (with integer coefficients) survives as full quantum duality symmetry of string theory.

Included in this $Sl(2,\mathbb{Z})$ transformations is the important transformation with matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \in Sl(2,\mathbb{Z}) , \tag{21.42}$$

which gives the transformation

$$\tau' = -\frac{1}{\tau} . \tag{21.43}$$

If $C_0 = 0$, this transformation becomes

$$\Phi' = -\Phi \Rightarrow g_s' = \frac{1}{g_s} , \tag{21.44}$$

which means that this is a nonperturbative duality that exchanges weak and strong coupling. Since now $\Lambda = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$, the transformation exchanges NS-NS and R-R fields,

$$H_3' = F_3, \quad F_3' = -H_3 . \tag{21.45}$$

The 5-form field is invariant $F_5' = F_5$, and also the Einstein metric is invariant.

21.4 M-Theory

It turns out that type IIA string theory at strong coupling, $g_s = e^{\langle\phi\rangle} \to \infty$, becomes 11 dimensional, and is called M-theory. This was proven by Ed Witten in a 1995 paper ("String theory in various dimensions"), in which it was also proven that various string theories are related by dualities, and thus in fact there is a single unified theory, as we will see at the end of the chapter.

We don't know too much about M-theory, in particular we don't have a quantum formulation, not even a perturbative one like in the case of string theory. The one thing that we know is about the low energy theory, i.e., in the limit of energies much

smaller than the 11 dimensional Planck scale, the theory can only be a supersymmetric field theory including gravity, so it must be the unique 11 dimensional supergravity. The theory has bosonic fields the metric G_{MN} and a 3-form antisymmetric tensor $A_{MNP} = A_{(3)}$, with field strength

$$F_{M_1 \ldots M_4} = 24 \partial_{[M_1} A_{M_2 M_3 M_4]} : \quad F_{(4)}. \tag{21.46}$$

The fermionic field is the gravitino $\psi_{\mu\alpha}$, but we will not need it in the following. We consider only the bosonic action, which is

$$S_{11}^{\text{bose}} = \frac{1}{2\kappa_{11}^2} \int d^{11}x \sqrt{-G} \left(R - \frac{1}{2} |F_4|^2 \right) - \frac{1}{6\kappa_{11}^2} \int d^{11}x A_{(3)} \wedge F_{(4)} \wedge F_{(4)}. \tag{21.47}$$

Under dimensional reduction, the theory must reduce to type II supergravity (the low energy limit of superstring theory). Specifically, since the 11 dimensional spinor splits into two 10 dimensional spinors of opposite chirality, it must be type IIA supergravity. The dimensional reduction ansatz for the metric is

$$\begin{aligned} ds_{11}^2 &= G_{MN}^{(11)} dx^M dx^N \\ &= e^{-\frac{2\Phi}{3}} G_{\mu\nu}^{s,(10)} dx^\mu dx^\nu + e^{\frac{4\Phi}{3}} (dx^{10} + A_\mu dx^\mu)^2 \,, \end{aligned} \tag{21.48}$$

and the reduction ansatz for the 3-form $A_{MNP} = (A_{\mu\nu\rho}, A_{\mu\nu 10})$ is that $A_{\mu\nu 10} = B_{\mu\nu}^{IIA}$ is the NS-NS B-field, whereas $A_{\mu\nu\rho} = A_{\mu\nu\rho}^{IIA}$ is the R-R 3-form, completing the bosonic fields of type IIA supergravity.

Relation Among Parameters

In the dimensional reduction metric, the coefficient of $(dx^{10})^2$ must be the radius squared in Planck units, $(R/l_P)^2$, but it is actually $e^{4\Phi/3}$. which means that

$$\left(\frac{R}{l_P} \right)^2 = e^{\frac{4\langle\Phi\rangle}{3}} = g_s^{4/3} \Rightarrow R = l_P g_s^{4/3}. \tag{21.49}$$

Here l_P is the 11 dimensional Planck length, so $2\kappa_{11}^2 = (2\pi)^8 l_P^9$. Dimensionally reducing the 11 dimensional Einstein–Hilbert term, we get

$$\frac{2\pi R}{2\kappa_{11}^2} \int d^{10}x e^{-2\Phi} R = \frac{1}{2\kappa_{10}^2} \int d^{10}x e^{-2(\Phi - \langle\Phi\rangle)} R \,, \tag{21.50}$$

where $2\kappa_{10}^2 = g_s^2 (2\pi)^7 (\sqrt{\alpha'})^8$. Identifying the two sides, we get

$$(2\pi)^7 l_P^9 = R g_s^2 (\sqrt{\alpha'})^8 (2\pi)^7 \Rightarrow l_P = g_s^{1/3} \sqrt{\alpha'}. \tag{21.51}$$

Combining with the previous relation, $R = l_P g_s^{2/3}$. we get

$$R = g_s \sqrt{\alpha'} \,, \tag{21.52}$$

i.e., the radius is the string coupling in string units, as we had advertised.

21.5 The String Duality Web

We have seen that under T-duality, Dp-branes change to D$(p+1)$-branes or D$(p-1)$-branes, and that type IIA string theory has even Dp-branes, whereas type IIB has odd Dp-branes. We can prove moreover that in fact the full type IIA string theory maps to type IIB string theory under T-duality. We also saw that type IIB string theory is invariant under S-duality.

Besides type II string theories, we also have type I and heterotic string theories, both of which reduce at low energies to $\mathcal{N} = 1$ supergravity, coupled to $\mathcal{N} = 1$ gauge theories for gauge groups $SO(32)$ (for both type I and heterotic) and $E_8 \times E_8$ (only for heterotic).

In fact, as we mentioned, it was proven by Witten that there is only one theory, and the various string theories and M-theory are connected to it, making a duality web. We try to describe here these dualities.

- S duality applied to the type I $SO(32)$ theory gives the $SO(32)$ heterotic theory, giving
$$\Phi_I = -\Phi_h \Rightarrow g_s^I = \frac{1}{g_s^{\text{heterotic}}}. \tag{21.53}$$
The Einstein metric is invariant, $G^{E,I}_{\mu\nu} = G^{E,h}_{\mu\nu}$, the gauge fields also, $A^I_{(1)} = A^h_{(1)}$, and the 3-forms are identified, $\tilde{F}^I_{(3)} = \tilde{H}^h_{(3)}$.
- T-duality applied to the $E_8 \times E_8$ heterotic string (on S^1) gives the $SO(32)$ heterotic string.
- T-duality of type I $SO(32)$ string theory on S^1 gives type IIA string theory on the interval $S^1/\mathbb{Z}_2$ (obtained by identifying the circle under $x \leftrightarrow -x$), which is M-theory on $S^1 \times (S^1/\mathbb{Z}_2)$.
- The strong coupling limit of the $E_8 \times E_8$ heterotic string is M-theory on the interval $S^1/\mathbb{Z}_2$, with "M9-branes" (10 dimensional walls in 11 dimensions, similar to the D-branes in string theory, having gauge fields on the worldvolume) at the endpoints. Each M9-brane carries a group E_8 on its worldvolume, for a total of $E_8 \times E_8$.

The relations above then connect all possible string theories, together with M-theory, in a "duality web", so that all are in fact corners in moduli space of the same (unknown) theory. Indeed, under S-duality, two theories are equivalent, but when one is perturbative, the other is nonperturbative, so only one is a good (calculable) description at any point. Similarly, under T-duality, among the equivalent theories only one (with $R > \sqrt{\alpha'}$) is a good perturbative description.

Important Concepts to Remember

- T-duality for open strings changes $X = X_L + X_R$ into $X' = X_L - X_R$, and the open string now stretches between two D-branes.
- Thus T-duality exchanges Neumann with Dirichlet boundary conditions, and turns a Dp-brane into a D$(p+1)$-brane (for T-duality on a transverse direction) or D$(p-1)$-brane (for T-duality on a parallel direction)
- T-duality with Chan–Paton factors, i.e., with nonabelian gauge fields, in the presence of Wilson lines, leads to D-branes in the nonabelian gauge group.
- For a Wilson loop breaking $U(N) \to U(1)^N$ by $A = \text{diag}(\theta_1, \ldots, \theta_N)$, the D-branes are at different positions $\theta_1, \ldots, \theta_N$.
- We thus have a quantum geometry, with "positions" for N D-branes that are now $N \times N$ matrices.
- Low energy superstring theory in 10 dimensions gives type II supergravity, specifically IIA if the spinors have opposite chiralities, and IIB if they have different chiralities.
- The field content of supergravity is the NS-NS sector, $(g_{\mu\nu}, B_{\mu\nu}, \phi)$, common with the bosonic string, the fermionic NS-R and R-NS sectors, and the antisymmetric tensors in the R-R sector: odd p-form potentials in type IIA, even in type IIB.
- The type IIB supergravity action can only be written covariantly (and without auxiliary fields) by imposing the self-duality constraint for the 5-form field strength, $F_5 = *F_5$, by hand.
- The IIB action has S-duality, $Sl(2, \mathbb{R})$ invariance of the equations of motion and $Sl(2, \mathbb{Z})$ restriction at the quantum level, acting on $\tau = C_0 + ie^{-\phi}$. In it, we have the transformation $\tau' = -1/\tau$ that inverts the string coupling and changes the RR with the NS-NS form fields.
- Type IIA string theory at strong coupling becomes 11 dimensional M-theory, with the 11th dimension being the string coupling, $R = g_s\sqrt{\alpha'}$, or $R = g_s^{2/3} l_P$.
- The 11 dimensional supergravity, the low energy limit of M-theory, KK reduces on a circle to the 10 dimensional type IIA supergravity.
- All string theories, type IIA, IIB, type I, heterotic $E_8 \times E_8$ and $SO(32)$, as well as M-theory, are related in a string duality web by T-duality, S-duality and dimensional reduction.

Further reading: See [21] for more details.

Exercises

(1) *(Buscher T-duality)*
Consider the string Polyakov action for $B_{\mu\nu} = 0$, $\phi = \phi_0$,

$$S_P = -\frac{1}{4\pi\alpha'} \int d^2\sigma \sqrt{-\gamma}\gamma^{ab} g_{\mu\nu} \partial_a X^\mu \partial_b X^\nu \ , \tag{21.54}$$

with $g_{\mu\nu}$ split into (g_{ij}, g_{i0}, g_{00}), and 0 is a T-duality direction.

Find a first order form for X^0 by replacing it with a $V_a = \partial_a X^0$ variable and adding a Lagrange multiplier. By solving for X^0, find the dual action in terms of the Lagrange multipliers $\hat{X}^0$.
(2) Check that under KK dimensional reduction of 11 dimensional supergravity, the terms involving $A_{(3)}$ reduce to the type IIA supergravity terms involving $B_{(2)}$ and $A_{(3)}$.
(3) Prove that "D0-branes" ("D-particles") in type IIA string theory, objects charged under the type IIA A_μ field, with masses $M_n = \frac{|n|}{g_s\sqrt{\alpha'}}$, $n \in \mathbb{Z}$, are momentum modes in the 11 dimensional supergravity on a circle.
(4) Show that both type IIB and type IIA supergravity bosonic actions give the same action by dimensional reduction on a circle to 9 dimensions.
(5) Consider D-branes situated at $\Delta\theta$ distance between them. When do we find massless modes for the strings between them, and when these modes are the ground state ("tachyons")?

Chapter 22
String Theory and Particle Physics

In this next to last chapter of part II of this book, we finally consider how to obtain particle physics from string theory. Indeed, string theory is supposed to describe the real world, and since moreover we want to describe cosmology, we need to check whether what we obtain is consistent with the particle physics that we know, i.e., the Standard Model.

–The standard way to obtain particle physics is via dimensional reduction on the compact space.

–The alternative is to consider a *braneworld* in the string theory context, i.e., the particle physics lives on a brane in some curved space, whereas gravity can live in the bulk of spacetime. In the context of string theory, the brane refers to a D-brane, or more precisely for the intersection of D-branes.

22.1 Dimensional Reduction on Compact Spaces

We start the analysis with the standard case, with KK dimensional reduction. In 10 dimensional superstring theory, we consider the standard set-up, $M_{10} = M_4 \times K_6$. Moreover, we want to obtain $\mathcal{N} = 1$ supersymmetry in 4 dimensions after compactification.

The reason is that in 4 dimensions it is very hard to break $\mathcal{N} > 1$ to $\mathcal{N} = 1$ only, and we need only $\mathcal{N} = 1$ for phenomenological reasons: $\mathcal{N} = 2$ and higher supermultiplets come in chiral pairs (non-complex representation), whereas in the real world we have chiral fermions. The only other possibility is no supersymmetry at all ($\mathcal{N} = 0$), but then, what is the point of string theory, if we don't even use supersymmetry to solve the standard problems of particle physics?

$\mathcal{N} = 1$ unbroken susy means that the vacuum $|\psi\rangle$ is invariant under the susy, i.e.,

$$Q|\psi\rangle = 0 \,, \tag{22.1}$$

which on the fields means that

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_22

$$\delta_Q \text{fields} = 0. \tag{22.2}$$

But since δ bosons = fermions, and fermions have zero VEV, in vacuum the variation of the bosons is automatically satisfied, and we only need to impose as a constraint

$$\delta_Q \text{fermions} = 0. \tag{22.3}$$

The most popular string theory for compactifications leading to Standard Model-like theories, and also the first to be well established, is the $\mathcal{N} = 1$ *heterotic superstring*, which in the supergravity limit has the fermion variations

$$\begin{aligned}
\delta\psi_M &= \frac{1}{\kappa_N} D_M \eta + \frac{\kappa_N}{2g^2\phi} \left(\Gamma_M{}^{NPQ} - 8\delta_M^N \Gamma^{PQ}\right) \eta H_{NPQ} + (\text{fermi})^2 \eta \\
\delta\chi^a &= -\frac{1}{4g\sqrt{\phi}} \Gamma^{MN} F^a_{MN} \eta + (\text{fermi})^2 \eta \\
\delta\lambda &= \frac{1}{\sqrt{\phi}} (\Gamma \cdot \partial\phi)\eta + \frac{\kappa_N}{6\sqrt{2}g^2\phi} \Gamma^{NPQ} \eta H_{NPQ} + (\text{fermi})^2 .
\end{aligned} \tag{22.4}$$

Here ψ_M is the gravitino (superpartner of the graviton), χ^a is the gluino (superpartner of the gauge fields) and λ is the dilatino, and $H = dB$ is the field strength of the NS-NS B-field. For supersymmetry, we need to put to zero all three variations (of the gravitino, gluino and dilatino). We have not written explicitly the fermion squared terms, since in the vacuum (as VEV) they must vanish.

The standard way to solve the vanishing of all 3 variations (but by no means the only one; recent models obtaining the Standard Model spectrum violate this simple case) is via $H = d\phi = 0$. In this case, we can easily see that the solution of the equations is

$$\begin{aligned}
&D_i \eta = 0 \\
&\Gamma^{ij} F_{ij} \eta = 0 \, ,
\end{aligned} \tag{22.5}$$

where i, j are indices in K_6, the compact space (on M_4, the equations are trivial in the vacuum). Since we want a *single* supersymmetry, we need to have a *single covariantly constant spinor* in K_6.

Holonomy and Calabi–Yau Spaces

We can define *parallel transport* of a spinor along a curve in curved spacetime by the condition of vanishing covariant derivative (covariantly constant spinor),

$$D_i \eta = 0 \Rightarrow \eta(x + dx) = (1 + \omega \cdot dx)\eta(x) \, , \tag{22.6}$$

which leads to the finite form

$$\begin{aligned}
\eta^\alpha &\to U^\alpha{}_\beta \eta^\beta \\
U(\gamma) &= P \exp \int_\gamma \omega \cdot dx .
\end{aligned} \tag{22.7}$$

Here $\omega \cdot dx = \omega_\mu dx^\mu = \omega_\mu^{ab} \frac{1}{4} \Gamma^{ab} dx^\mu$ is given by the spin connection.

On a closed path γ, we call U the *holonomy* for the spinor. Since $\omega_\mu^{ab} \in SO(n)$ (is in the Local Lorentz group), then *generically* $U \in SO(n)$. That means that the *holonomy group*, the group formed by all possible holonomies of the space, is $\subseteq SO(n)$.

We can define a *complex manifold* by the statement that there is a matrix $J^i{}_j$ such that $J^2 = -\mathbb{1}$ (this technically means an almost complex manifold), plus an extra condition (involving the so-called Nijenhuis tensor). The statement is that locally, around a point, we can define the notion of i, and thus write complex coordinates.

We can further define a *Kähler manifold*, as the case that the holonomy group is $\subseteq U(N)$, where $N = n/2$ is the complex dimension of the manifold. In that case, $J^i{}_j$ is covariantly constant. An equivalent definition of the Kähler manifold is as a complex manifold where we can locally define a function K, called the Kähler potential, such that

$$g_{i\bar{j}} = \partial_i \partial_{\bar{j}} K. \tag{22.8}$$

In that case, after dimensional reduction on the manifold, we will obtain $\mathcal{N} = 1$ supersymmetry, where we have the same Kähler potential for the scalars corresponding to the compact coordinates, so the name is not a coincidence.

In the case of a Kähler manifold, the spin connection is in $U(N) \simeq SU(N) \times U(1)$ (modulo topological issues). But if the holonomy group is actually $SU(N)$, it means that the $U(1)$ part is topologically trivial (pure gauge, with zero field strength), or mathematically we say that the *first Chern class of the manifold K is zero*, $c_1(K) = 0$.

But Calabi and Yau proved a theorem saying that having a Kähler manifold with vanishing first Chern class $c_1(K) = 0$ is $\Leftrightarrow$ having a unique Kähler metric of $SU(N)$ holonomy. Such a space is then called a *Calabi–Yau manifold*.

On a Calabi–Yau manifold, there is a unique covariantly constant spinor η, $D_i \eta = 0$. But then,

$$[D_i, D_{\bar{j}}]\eta = 0 \Rightarrow R_{i\bar{j}} = 0 \,, \tag{22.9}$$

so the manifold is Ricci-flat (the commutator of two covariant derivatives is proportional to the field strength, and for ω the field strength $R_{ij}(\omega)$ is proportional to the Ricci tensor).

So if we take a Calabi–Yau space when compactifying string theory, we not only get a unique covariantly constant spinor, guaranteeing $\mathcal{N} = 1$ supersymmetry in 4 dimensions, but we also satisfy the Einstein equation in the compact space, since $R_{i\bar{j}} = 0$.

If η is the unique covariantly constant spinor, then

$$k_{ij} = \bar{\eta} \Gamma_{ij} \eta \tag{22.10}$$

is the Kähler form, such that when contracted with the metric it gives the complex structure,

$$J^i{}_j = g^{ik}k_{kj}. \tag{22.11}$$

If moreover we consider a 3-dimensional Calabi–Yau CY_3, we also have a *holomorphic 3-form*

$$\Omega_{ijk} = \bar{\eta}\Gamma_{ijk}\eta. \tag{22.12}$$

On CY_n, we will have a holomorphic n-form.

Moduli Space of CY_3

A modulus is a scalar that has no perturbative potential, so it costs no energy to change it. In the case of KK compactifications, parameters that don't have a potential become moduli. For CY_3, we ask also that they don't change the CY_3 structure; then, we have a number of such deformations corresponding to moduli, that fall within two categories:

–*complex structure moduli* are moduli than deform the shape of the space. The simplest example of a Calabi–Yau space is an even torus, and the simplest example of that is T^2. In this case, there is one complex structure modulus, the ratio of radii, R_2/R_1. More precisely, since the modulus needs to be a complex, the "ratio" includes an angle for the periodicities of the identified directions on the plane that give the torus, and is the modular parameter τ of the torus.

–*Kähler structure moduli* are moduli that deform the size of the space. In the case of T^2, the only size modulus is the volume, R_1R_2, or more precisely a complexified version that includes a B-field on the torus.

The moduli (deformations of the CY, which in particular is Kähler) will form a scalar space with no (perturbative) potential, but with nontrivial kinetic term. In fact, they will also form a Kähler space, which is why their kinetic term is written in terms of a Kähler potential, $g_{i\bar{j}}\partial_\mu\phi^i\partial^\mu\phi^{\bar{j}}$, with $g_{i\bar{j}} = \partial_i\partial_{\bar{j}}K$.

Moreover, it is an important result, due to Strominger, that while the Calabi–Yau compactification preservers $\mathcal{N} = 1$ supersymmetry, the moduli preserve $\mathcal{N} = 2$ supersymmetry, and they fall into $\mathcal{N} = 2$ multiplets, in particular $\mathcal{N} = 2$ vector multiplets (made up of $\mathcal{N} = 1$ scalar (WZ) and vector multiplets) and $\mathcal{N} = 2$ hypermultiplets (made up of two $\mathcal{N} = 1$ scalar, or WZ, multiplets).

22.2 Calabi–Yau Compactification of String Theories

Type IIB String Theory on CY_3

In this case, the low energy theory has the following $\mathcal{N} = 2$ supermultiplets: the $\mathcal{N} = 2$ supergravity multiplet, $n_v = b_3$ (here b_3 is the "third Betti number", a topological quantity that characterizes the manifold) vector multiplets containing the complex structure moduli, and $n_h = b_2 + 1$ hypermultiplets (b_2 is the second Betti number, also a topological number) containing Kähler structure moduli. There is

also an extra complex structure modulus, the complex scalar τ made up of the axion a and the dilaton ϕ by

$$\tau = a + ie^{-\phi}. \tag{22.13}$$

Type IIB string theory by itself is not phenomenologically interesting, and the interest in it only arose once people started to introduce *G-fluxes*, which mean the nonzero integral over a cycle in the geometry of an antisymmetric tensor G defined as

$$G = F^{\mathrm{RR}} - \tau H^{\mathrm{NS-NS}}. \tag{22.14}$$

Here H is the field strength of the NS-NS B-field, and F is the field strength of the RR 2-form. The case of theories with G-flux was defined by Giddings, Kachru and Polchinski, and then Gukov, Vafa and Witten calculated the resulting superpotential in 4 dimensions, and found it to be

$$W = \int_{K_6} \Omega \wedge G \ , \tag{22.15}$$

where Ω is the holomorphic 3-form.

Always there is at least one Kähler modulus, the complexified volume ρ. We also have at least a complex structure τ, as we saw, and the rest are called τ_α. In this case, the Kähler potential is

$$\begin{aligned} K(\rho, \tau, \tau^\alpha) &= K(\rho) + K(\tau, \tau^\alpha) \\ K(\rho) &= -3\ln[-i(\rho - \bar{\rho})] \\ K(\tau, \tau^\alpha) &= -\ln[-i(\tau - \bar{\tau})] - \ln\left[-i\int_{K_6} \Omega \wedge \bar{\Omega}\right]. \end{aligned} \tag{22.16}$$

Heterotic $E_8 \times E_8$ String Theory on CY_3

The first compactification model to become popular was the heterotic $E_8 \times E_8$ model, specifically on CY_3. We have seen the general conditions for $\mathcal{N} = 1$ susy in 4 dimensions. But the NS-NS field strength has an anomalous Bianchi identity. Instead of $dH = 0$ (since $H = dB$), in the heterotic $E_8 \times E_8$ model we have in fact

$$dH = \mathrm{Tr}(F \wedge F) - \mathrm{Tr}(R \wedge R) \ , \tag{22.17}$$

where in the first term, the trace refers to the $E_8 \times E_8$ YM group, for F_{ij}^{AB}, and in the second term, the trace refers to the local Lorentz group, for R_{ij}^{ab}. We have seen that the simplest model we considered had $H = 0$, which means from the above that we must have $\mathrm{Tr}(F \wedge F) = \mathrm{Tr}(R \wedge R)$. The simplest way to satisfy this is using "$F = R$". Specifically, this can be done by "embedding the spin connection into the gauge group," i.e., identifying the 4 dimensional fields coming from the gauge field A_i and the metric g_{ij} (or more precisely the spin connection ω_i^{ab}), by (here A, B are $SO(16) \subset E_8$ indices)

$$A_i^{AB} = \begin{pmatrix} 0 & 0 \\ 0 & \omega_i^{ab} \end{pmatrix}. \tag{22.18}$$

This is a common thing in KK compactifications, as we said: the various fields with the same index structure (same spin) mix up, and thus here the two fields in A_i and g_{ij} have the same lower dimensional field in their ansatz. For this embedding, the gauge group is broken to the subgroup of E_8 commuting with the ansatz (i.e., with the $SO(6)$ of the spin connection), namely $SO(10)$.

One can also consider nontrivial Wilson loops in the geometry, i.e.,

$$U_\gamma \equiv P \exp \oint_\gamma A \cdot dx \neq 0 \tag{22.19}$$

for some γ. As we said also in the previous chapter, this breaks the gauge group further (for a general Wilson line, this would break $U(N)$ to $U(1)^N$, but in specific cases, we can break it less), towards the Standard Model group, $SU(3) \times SU(2) \times U(1)$.

This concludes our analysis of KK compactification models. There is also the possibility of string theory near some special singularities, but we will not study that. Then there is only the idea of braneworld.

22.3 Intersecting D-Branes: Braneworld

In this construction, within string theory, the Standard Model of particle physics is supposed to live on the intersection of D-branes, whereas gravity exists everywhere. This seems to have a chance of working, since the Standard Model gauge group is $SU(3) \times SU(2) \times U(1)$, and on N coincident branes we have an $U(N) \simeq SU(N) \times U(1)$ group, where the $U(1)$ multiplet contains the scalar parametrizing the center of mass motion of the N D-branes.

Since we want to obtain the Standard Model in 3+1 dimensions on the intersection of branes, we need to have Dp-branes with $p > 3$. In fact, it turns out that the most flexible case is $p = 6$, since then we have more possibilities. Thus the intersecting D6-branes models are the most popular.

As a simple exercise (not even remotely phenomenologically correct), we consider D6-brane on the torus T^6, and even more specifically, we consider two D6-branes, D6 and D6'. In reality, we will have some discrete identification of T^6 directions (using maybe $\mathbb{Z}_n$ identifications), and usually more D-branes, for a more realistic case. Splitting the T^6 into 3 tori for coordinates (x_4, x_5), (x_6, x_7) and (x_8, x_9), we consider the directions filled by D6 and D6' to be denoted with a cross,

	0	1	2	3	4	5	6	7	8	9
$D6$	×	×	×	×	×	0	×	0	×	0
$D6'$	×	×	×	×	0	×	0	×	0	×

(22.20)

and the D6 and D6' intersect in each T^2 in the middle of the square torus.

More generally, we can consider intersections of D6-branes at angles, defined by some integers. For instance, the D-brane with $l = (3, 1)$ is identified over 3 fundamental domains in the x direction and one domain in the y direction.

When D-branes intersect, there are extra massless open string states that appear at the intersection. Indeed, when we stretch open strings between D-branes i and j we have a mass $M_{ij} = \frac{1}{2\pi\alpha'} r_{ij}$, so if the distance $r_{ij} = 0$, also $M_{ij} = 0$. Thus indeed, at the intersection (and only there), where $r_{ij} = 0$ between the intersecting branes, we have massless states. However, for D-branes at angles different than $\pi/2$, the bifundamental scalars coming from r_{ij} are generically massive, whereas the bifundamental fermions are always massless. If $i \in SU(m)$ (for m coincident branes on the left) and $j \in SU(q)$ (for q coincident branes on the right), the state (ij) is bifundamental in $SU(m) \times SU(q)$. If for instance, we manage to "freeze" the $SU(q)$ gauge fields, so that $SU(q)$ becomes a global symmetry and $SU(p)$ remains local, then this matter is really like the quark matter in the Standard Model, with one local and one global index.

22.4 Moduli Stabilization

We have seen that in string theory we have many moduli, i.e., scalars that are massless or nearly massless perturbatively. But experimentally, we know there aren't any, at least on the scales of the Earth, since otherwise we would detect a fifth force. There could be such scalars on cosmological scales, hidden from Earth experiments, in the so called chameleon theories, that will be studied later on, but certainly there aren't on Earth scales.

That means that we need to "stabilize" these moduli, by introducing some nonperturbative potentials. But in string theory, that is obviously difficult, as we will discuss in part III of the book, since we only have a perturbative definition of string theory, and nonperturbative quantities we can compute are few, like ones found through dualities, or nonperturbative contributions to the superpotential coming from wrapped branes (since the D-brane action is proportional to $1/g_s$, by having terms of the e^{iS} type, we get nonperturbative contributions).

The moduli we need to stabilize are:

- shape and size moduli, i.e., complex and Kähler structure moduli.
- the dilaton is a special case, since it defines the string coupling constant, and is usually considered separately.
- if we have D-branes, they come with their own moduli, for instance for their positions.
- if we have fluxes, the fluxes usually have integers values, but sometimes they also lead to moduli.

Generically, it is very hard to stabilize *all* the moduli. Usually, the last modulus to stabilize turns out to be difficult to do so.

Important Concepts to Remember

- Particle physics is obtained from string theory either by KK dimensional reduction, or in a braneworld scenario.
- For phenomenological reasons we need $\mathcal{N} = 1$ supersymmetry in 4 dimensions, which means that we need to put to zero the supersymmetry variation of the fermions in the vacuum, $\delta\psi = 0$.
- For heterotic supergravity, the standard way is to use $H = d\phi = 0$, giving $D_i\eta = 0$ and $F_{ij}\Gamma^{ij}\eta = 0$.
- A Kähler manifold is a complex manifold that has holonomy inside $U(N)$.
- If the holonomy is actually $SU(N)$, so $c_1(K) = 0$ (first Chern class of K is zero), then there is a unique Kähler metric of $SU(N)$ holonomy, and the space is called Calabi–Yau manifold.
- On a Calabi–Yau manifold, there is a unique covariantly constant spinor, $D_i\eta = 0$, so we obtain $\mathcal{N} = 1$ supersymmetry in 4 dimensions, and $R_{i\bar{j}} = 0$, so the manifold is Ricci-flat.
- Calabi–Yau spaces have complex structure moduli, that deform the shape, and Kähler structure, that deform the size.
- For type IIB on CY_3 we can introduce G-flux, for $G = F^{\mathrm{R-R}} - \tau H^{\mathrm{NS-NS}}$, with superpotential $W = \int \Omega \wedge G$, where Ω is the unique holomorphic 3-form on CY_3.
- For heterotic $E_8 \times E_8$ string theory on CY_3, we can embed the spin connection in the gauge group, and put Wilson lines, to break the gauge group to $SU(3) \times SU(2) \times U(1)$ or a GUT group.
- A braneworld model can be obtained for instance by intersecting D6-branes in the compact space.
- The moduli must be stabilized for phenomenological reasons.

Further reading: See [20–22] for more details.

Exercises

(1) If $H_{MNP} = R\epsilon_{MNP}$, and we are on a sphere, write down the condition for the spinor η that preserves susy.
(2) Assuming that (in type IIB on CY_3)

$$\int_{K_6} \Omega \wedge \bar{\Omega} = 0 \ , \quad \int_{K_6} \Omega \wedge G = \rho e^{a\rho} \ , \tag{22.21}$$

calculate the scalar potential (the above is not a very realistic case).
(3) In the model with D6, D6'-brane on T^6, calculate the open string spectrum on the common intersection.
(4) Show that by "embedding the spin connection in the gauge group", which as we said leads to the group $SO(10)$, we can neatly fit the Standard Model particles, plus a right-handed neutrino, into a single representation of the unbroken gauge group.
(5) Is a sphere a Calabi–Yau space? Why?

Chapter 23
Holography and the AdS/CFT Correspondence

In this chapter, we will describe the AdS/CFT correspondence, which relates a gravity theory, usually in an Anti-de Sitter (AdS) $d+1$ dimensional background to a field theory with conformal invariance (CFT) in d dimensions, living at a boundary of the space. Therefore the correspondence is an example of a holographic relation, the gravitational physics in the bulk being described holographically ("projected" without losing information) onto the non-gravitational boundary of one dimension less. The AdS/CFT correspondence will be used later on, in part III, for constructing cosmological models. In order to understand that, we will describe the AdS/CFT map for fields and observables.

23.1 AdS Space and Its Holography

In cosmology, the relevant gravitational spacetimes, besides flat space and the generic FLRW metric, are de Sitter space, approximated during the inflationary period, and its deformations, perhaps relevant for today's accelerated expansion, in the case of a cosmological constant. The cosmological constant, both for inflation, and for today's accelerated expansion, is positive.

However, one could consider also a negative cosmological constant, $\Lambda < 0$, leading to an Anti-de Sitter space. This would not be relevant as a cosmological model, but it turns out to be relevant in string theory, and to lead to an application of string theory not as a fundamental theory of Nature, but as a mathematical tool to solve difficult, i.e., strong coupling, field theory problems.

Anti-de Sitter space in d dimensions (AdS_d) is a maximally symmetric, homogenous and isotropic, gravitational space of constant negative curvature. In Euclidean signature, the maximally symmetric spaces, besides flat space, are the sphere S^d, which is a space of constant positive curvature $R > 0$, and the Lobachevsky space, which is a space of constant negative curvature $R < 0$.

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_23

The sphere is defined via embedding in $d+1$ dimensional Euclidean space

$$ds^2 = dX_1^2 + \cdots + dX_d^2 + dX_{d+1}^2 \tag{23.1}$$

by the constraint with the same $SO(d+1)$ symmetry,

$$(X_1)^2 + \cdots (X_d)^2 + (X_{d+1})^2 = R^2. \tag{23.2}$$

Lobachevsky space is defined similarly, but by embedding in $d+1$ dimensional Minkowski space (despite having Euclidean signature)

$$ds^2 = dX_1^2 + \cdots + dX_d^2 - dX_{d+1}^2 \tag{23.3}$$

by the constraint with the same $SO(d,1)$ symmetry,

$$(X_1)^2 + \cdots (X_d)^2 - (X_{d+1})^2 = -R^2. \tag{23.4}$$

Moving on to Minkowski signature spaces, the sphere becomes de Sitter space, as the space of constant positive curvature (so constant cosmological constant), defined by embedding in $d+1$ dimensional Minkowski space

$$ds^2 = -dX_0^2 + dX_1^2 + \cdots dX_{d-1}^2 + dX_{d+1}^2 \tag{23.5}$$

by the constraint with the same $SO(d,1)$ symmetry,

$$-(X_0)^2 + (X_1)^2 + \cdots (X_{d-1})^2 + (X_{d+1})^2 = R^2. \tag{23.6}$$

Thus the isometry of de Sitter is $SO(d,1)$, and "Euclidean de Sitter" is the sphere, $EdS_d = S^d$.

Then Lobachevsky space becomes Anti-de Sitter, as the space of constant negative curvature (so negative cosmological constant), defined by embedding in a $d+1$ dimensional space with two "times",

$$ds^2 = -dX_0^2 + dX_1^2 + \cdots dX_{d-1}^2 - dX_{d+1}^2 \tag{23.7}$$

by the constraint with the same $SO(d-1,2)$ symmetry,

$$-(X_0)^2 + (X_1)^2 + \cdots (X_{d-1})^2 - (X_{d+1})^2 = -R^2. \tag{23.8}$$

That means that AdS space has $SO(d-1,2)$ isometry, and "Euclidean Anti-de Sitter" $EAdS_d$ is Lobachevsky space.

Besides the implicit metric obtained by solving, say for X_{d+1} in terms of the others, there are other forms of the AdS metric. One relevant one is the Poincaré metric,

$$ds^2 = \frac{R^2}{x_0^2}\left(-dt^2 + \sum_{i=1}^{d-2} dx_i^2 + dx_0^2\right). \tag{23.9}$$

Here $t, x_i \in \mathbb{R}$ and $0 < x_0 < +\infty$, yet this metric only covers part of AdS space, the "Poincaré patch". Another form of the Poincaré metric is obtained by $x_0/R = e^{-y}$, giving

$$ds^2 = e^{2y}\left(-dt^2 + \sum_{i=1}^{d-2} dx_i^2\right) + R^2 dy^2. \tag{23.10}$$

The boundary of the space (or rather, of the Poincaré patch) is at $y \to \infty$, i.e., at $x_0 = 0$. But it takes a finite time for light (moving on $ds^2 = 0$) to reach it from a finite point, since then

$$t = R\int_{y_0}^{\infty} e^{-y} dy = Re^{-y_0} < \infty. \tag{23.11}$$

That means that the boundary of the Poincaré patch *is a finite time away*.

Coordinates that cover the whole of AdS space are the "global coordinates", with metric

$$ds^2 = R^2(-\cosh^2\rho\, d\tau^2 + d\rho^2 + \sinh^2\rho d\Omega_{d-2}^2). \tag{23.12}$$

The further coordinate transformation $\tan\theta = \sinh\rho$ leads to another form for the global coordinate metric,

$$ds^2 = \frac{R^2}{\cos^2\theta}(-d\tau^2 + d\theta^2 + \sin^2\theta d\Omega_{d-2}^2)\ , \tag{23.13}$$

and the range of the coordinates is $0 \le \theta \le \pi/2$ and arbitrary τ.

To understand this space, we construct the Penrose diagram, which encodes causal and topological information. For it, we can drop conformal factors, obtaining Einstein's static Universe

$$ds^2 = -d\tau^2 + d\theta^2 + \sin^2\theta d\Omega_{d-2}^2. \tag{23.14}$$

Thus the Penrose diagram is an infinite strip, $\mathbb{R}_\tau \times [0, \pi/2]$, rotated around the sphere directions, giving a cylinder. The Poincaré patch is just a triangle region of the infinite strip, with the base at $\theta = 0$, see Fig. 23.1.

In this form, it is obvious that the boundary of space is the surface of the cylinder, at $\theta = \pi/2$,

$$ds^2 = -d\tau^2 + d\Omega_{d-2}^2. \tag{23.15}$$

The time it takes light to go from a finite point at $\theta = \theta_0$ to the boundary of space $\theta = \pi/2$ in global coordinates is now again finite,

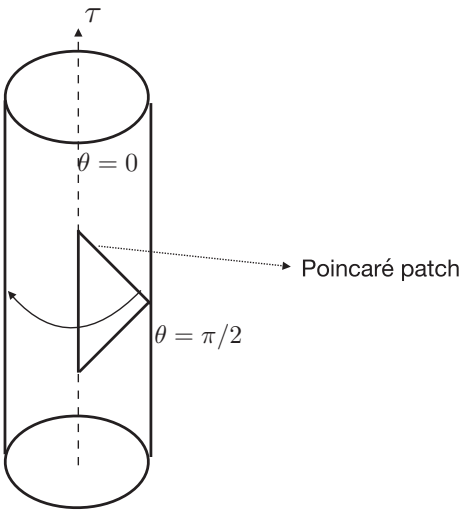

Fig. 23.1 Penrose diagram for global AdS, with the triangle corresponding to the Poincaré patch on it. The diagram is a cylinder of revolution

$$\tau = \frac{\pi}{2} - \theta_0. \tag{23.16}$$

Since the boundary of AdS_d space is a finite time away, light can go the boundary and back in finite time, which means that logically there is the possibility that we can project holographically the physics in the bulk of the space onto the boundary, which is $\mathbb{R}_\tau \times S^{d-2}$ in global coordinates and $\mathbb{R}^4$ in Poincaré coordinates. These two spaces are related by a conformal factor in Euclidean signature, since

$$ds^2 = \sum_{i=1}^{d} (dx_i)^2 = d\tilde{\rho}^2 + \tilde{\rho}^2 d\Omega_{d-1}^2 = \tilde{\rho}^2 \left(\frac{d\tilde{\rho}^2}{\tilde{\rho}^2} + \Omega_{d-1}^2 \right) = e^{-2y} (dy^2 + d\Omega_{d-1}^2) \,, \tag{23.17}$$

where in the last step we wrote $\tilde{\rho} = e^{-y}$.

Moreover, we saw that the isometry group of AdS_d space is $SO(d-1,2)$, and it will turn out that this is the conformal symmetry group in $d-1$ dimensions, suggesting that the holographic theory on the boundary of AdS space is a conformal field theory.

23.2 Conformal Field Theories

To understand the relation, we need to understand conformal field theories. Conformal transformations are generalizations of scale transformations of flat space, $x'^\mu = \lambda x^\mu$, which relate metrics as

$$ds^2 = (dx'^\mu)^2 = \lambda^2 (dx^\mu)^2. \tag{23.18}$$

For conformal transformations, the constant conformal factor λ^2 above is replaced by a local conformal factor. Thus conformal transformations are coordinate transformations $x^\mu \to x'^\mu(x^\nu)$ of flat space, such that the flat space metric gets a local conformal factor

$$ds^2 = (dx'^\mu)^2 = [\Omega(x)]^2 (dx^\mu)^2. \tag{23.19}$$

It turns out that most, if not all, theories with scale invariance at the quantum level have also conformal invariance. In fact, a characteristic of conformal invariant theories is that there are no parameters with mass scales in them (they are massless theories). The other observation is that conformal invariance is an invariance of flat space. So, even though a conformal transformation is part of general coordinate transformations, conformal invariance means that the flat space action, after the conformal transformation, is still a flat space action (it is independent of the conformal factor, so it is still written with the flat space metric).

Conformal field theories in $d = 2$ spacetime dimensions are special, since in 2 dimensions, any holomorphic transformation, $z \to z' = f(z)$, where $z = x_1 + i x_2$, is conformal:

$$ds^2 = dz' d\bar{z}' = |f'(z)|^2 dz d\bar{z}. \tag{23.20}$$

That amounts to an infinitely dimensional conformal group of transformations, which highly restricts possible theories.

But we will be interested in dimensions other than 2, $d > 2$, in which case the conformal group (group of conformal transformations) is finite dimensional. For an infinitesimal transformation $x'^\mu = x^\mu + v^\mu(x)$, with an infinitesimal $\Omega(x) \simeq 1 - \sigma_v(x)$, we obtain

$$\partial_\mu v_\nu + \partial_\nu v_\mu = 2\sigma_v \delta_{\mu\nu} \Rightarrow \sigma_v(x) = \partial^\mu v_\mu / d. \tag{23.21}$$

In $d > 2$, the most general solution of this equation is

$$v_\mu(x) = a_\mu + \omega_{\mu\nu} x_\nu + \lambda x_\mu + b_\mu x^2 - 2x_\mu b \cdot x \ , \tag{23.22}$$

giving $\sigma_v(x) = \lambda - 2b \cdot x$. Here λ is a scale transformation, corresponding to a dilatation generator D, a_μ is a translation, corresponding to the generator P_μ, the antisymmetric $\omega_{\mu\nu}$ corresponding to $J_{\mu\nu}$, and b_μ corresponds to a new type of transformation, a "special conformal transformation", for a generator K_μ. Together, the parameters fit into an antisymmetric matrix Ω_{MN}, with $M = (\mu, d+1, d+2)$, corresponding to generators of $SO(d, 2)$

$$\bar{J}_{MN} = \begin{pmatrix} J_{\mu\nu} & \bar{J}_{\mu,d+1} & \bar{J}_{\mu,d+2} \\ -\bar{J}_{\mu,d+1} & 0 & D \\ -\bar{J}_{\mu,d+2} & -D & 0 \end{pmatrix}. \tag{23.23}$$

Here $D = \bar{J}_{d+1,d+2}$ and

$$\bar{J}_{\mu,d+1} = \frac{K_\mu - P_\mu}{2}\,, \quad \bar{J}_{\mu,d+2} = \frac{K_\mu + P_\mu}{2}. \tag{23.24}$$

Thus indeed, as we have already reported, the isometry of AdS_{d+1} is the same as the conformal group in d flat spacetime dimensions, $SO(d, 2)$.

23.3 AdS/CFT Motivation (Heuristic Derivation)

The AdS/CFT correspondence was "motivated" (heuristically derived) by Maldacena in 1997 based on equating two descriptions of a large number of D-branes, in a certain decoupling limit that we will define.

We consider a near-extremal (i.e., only slightly non-extremal, mass only slightly larger than the charge) system of (a large number) N D-branes. As it was proven by comparing thermodynamic quantities, the temperature and entropy, given a mass and charge, first by Strominger and Vafa, then by Callan and Maldacena (both in early 1996), this system of D-branes describe the gravitational system of a near-extremal black hole.

But then in late 1996 Maldacena and Strominger go further, and prove explicitly that the Hawking radiation spectrum obtained from the near-extremal black hole matches the emission spectrum of the near-extremal large N system of D-branes, which means that at least in this toy model case, the Hawking radiation is unitary, specifically coming from a unitary string process happening on the D-branes. We can therefore understand this process of Hawking radiation from string theory as follows: two open strings, ending on a D-brane, collide, form a closed string, which can then move off the D-brane, as in Fig. 23.2. The closed string mode is a particle radiated by the black hole.

Then in his famous 1997 paper, Maldacena realized that by thinking of the two descriptions of this Hawking radiation, one gravitational, the other one in terms of D-branes, and taking a decoupling limit, we find the equivalence of a field theory living on the D-brane with a gravitational theory living near it. We present here the two points of view, and the equivalence resulting from it.

Point of View nr. 1: D-Branes, i.e., Open-String Endpoints

Thinking of Hawking radiation of the near-extremal D-branes as an open string process, the action describing the system is

$$S = S_{\text{bulk}} + S_{\text{brane}} + S_{\text{interaction}}. \tag{23.25}$$

On the brane we have open strings, in the bulk we have closed strings, and the two are coupled by an interaction that gives the Hawking radiation decay process. This interaction (for a closed string to go into two open strings) is of order g_s, the closed

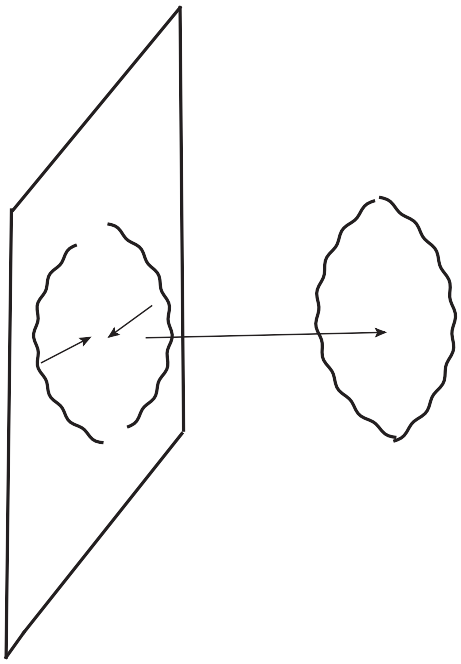

Fig. 23.2 Hawking radiation as a string process on a D-brane: two open strings on the D-brane join and move off the D-brane as a closed string, giving Hawking radiation

string coupling constant. But this is a gravitational interaction, proportional to the 10 dimensional Newton constant κ_N, which is proportional to g_s and to α'^2 (since in 10 dimensions $2\kappa_N^2$ has dimension -8, and α' has dimension -2). But that means that the interaction vanishes, in the "decoupling limit" $\alpha' \to 0$ (which means $\alpha' E^2 \to 0$ for relevant energies E), when moreover $\kappa_N \to 0$, so we obtain free (super)gravity, and also the D-brane action becomes just $\mathcal{N} = 4$ SYM.

Thus in this point of view, in the decoupling limit $\alpha' E^2 \to 0$ (where E is some relevant energy in the field theory) we have:

- free 10 dimensional gravity in the bulk, far from the branes (at $\delta r \to \infty$).
- $\mathcal{N} = 4$ SYM in 4 dimensions on the boundary.

Point of View nr. 2: (Near-)Extremal *p*-Brane Solutions, i.e., Black Holes

Now, replacing the N D-branes with the gravitational space curved by them, namely the type IIB supergravity solution

$$\begin{aligned} ds^2 &= H^{-1/2}(r) d\vec{x}_{||}^2 + H^{1/2}(r)(dr^2 + r^2 d\Omega_5^2) \\ F_5 &= (1+*)(dt \wedge dx_1 \wedge dx_2 \wedge dx_3 \wedge (dH^{-1})) \\ H(r) &= 1 + \frac{R^4}{r^4} \\ R &= 4\pi g_s N \alpha'^2 \ , \quad Q = g_s N \ , \end{aligned} \tag{23.26}$$

we have only a gravitational system, but it contains several regions. We consider the decoupling limit $\alpha' E^2 \to 0$, or $E\sqrt{\alpha'} \to 0$ but, since we are in gravitational theory, we need to keep the energy at a point p in the bulk, of radius r, fixed. Considering E as the energy measured at $r \to \infty$, the two are related as

$$E_p \sim \frac{d}{d\tau} = \frac{1}{\sqrt{-g_{00}}}\frac{d}{dt} \sim \frac{E}{\sqrt{-g_{00}}} , \tag{23.27}$$

which, since at $r \to 0$, $H(r) \sim R^4/r^4 \propto \alpha'^2/r^4$, means

$$\sqrt{\alpha'}E \sim (\sqrt{\alpha'}H^{-1/4})E_p \sim rE_p \to 0 \tag{23.28}$$

happens for $r \to 0$ if E_p is fixed.

So the decoupling limit is a limit of low energy at infinity, and then at infinity (for $\delta r \to \infty$) we get free gravity (since $\kappa_N \propto \alpha'^2 \to 0$, as before), but now also at $r \to 0$ we get $E \to 0$. That means that now we also have two decoupled systems:

- free gravity in the bulk at $\delta r \to \infty$
- low energy gravity excitations for $r \to 0$.

AdS/CFT Correspondence in Decoupling Limit

Identifying the two free systems in the two points of view, we see that one is the same (free gravity at $\delta r \to \infty$), which means the others must be identified:

$\mathcal{N} = 4$ **SYM with gauge group** $SU(N)$ **= gravitational theory in the D3-brane background at** $r \to 0$ and $\alpha' \to 0$.

We now interpret (23.28) at $r \to 0$ as

$$E \propto \frac{(E_p\sqrt{\alpha'})r}{\alpha'} \tag{23.29}$$

and, now identifying E with the energy of the SYM theory instead of the energy at infinity in the gravity description, we consider that, for the purposes of the comparison of SYM with gravity, we must keep fixed both the SYM energy E and the energy in the gravity (string) description in string units, so $\sqrt{\alpha'}E_p$, leading to

$$U = \frac{r}{\alpha'} \tag{23.30}$$

fixed in the $r \to 0$, $\alpha' \to 0$ limit.

In this limit, $H \sim R^4/r^4$, so we obtain

$$ds^2 \simeq \frac{r^2}{R^2}(-dt^2 + d\vec{x}_3^2) + \frac{R^2}{r^2}dr^2 + R^2 d\Omega_5^2 , \tag{23.31}$$

or, with the change of variables $r/R = R/x_0$,

$$ds^2 = R^2 \frac{-dt^2 + d\vec{x}_3^2 + dx_0^2}{x_0^2} + R^2 d\Omega_5^2 \,, \tag{23.32}$$

corresponding to the space $AdS_5 \times S^5$ of radius R. Writing it in terms of the variable $U = r/\alpha' \propto E/(E_p\sqrt{\alpha'})$, thus identified with the energy scale in SYM given a fixed energy in gravity, we obtain

$$ds^2 = \alpha' \left[\frac{U^2}{\sqrt{4\pi g_s N}}(-dt^2 + d\vec{x}_3^2) + \sqrt{4\pi g_s N}\left(\frac{dU^2}{U^2} + d\Omega_5^2\right)\right]. \tag{23.33}$$

In this form, there is a vanishing overall factor α', but the rest is finite. One can show that the string coupling g_s is related to the SYM coupling g_{YM} by

$$4\pi g_s = g_{YM}^2. \tag{23.34}$$

The proportionality comes becomes two basic open string interactions (proportional to g_{open}, i.e., g_{YM}) can be glued into a basic closed string interaction, or "pair of pants" (proportional to g_{closed}, i.e., g_s), and the proportionality constant can be calculated.

In this way, the $AdS_5 \times S^5$ metric is written in terms of SYM variables N, g_{YM} and U. The SYM field theory in some sense lives at the boundary of $AdS_5 \times S^5$, which is $\mathbb{R}^4$ (the S^5 shrinks to zero when going to the boundary of AdS_5), and happens as we saw for $x_0 \to 0$, or $r \to \infty$. Thus the AdS/CFT correspondence is *holographic*.

23.4 AdS/CFT Definition and Limits

We have obtained the AdS/CFT correspondence as an equality between $\mathcal{N} = 4$ SYM and string theory in the $AdS_5 \times S^5$ background given in (23.33), where N is either the rank of $SU(N)$, or the brane charge, and $4\pi g_s = g_{YM}^2$.

One could say that is all, and that is in fact the most general understanding of the AdS/CFT correspondence, currently believed to be correct. But as things stand, there are some approximations that put the heuristic derivation above on a stronger footing. Namely, we should be able to use the gravitational limit of string theory, which means that α' corrections and g_s corrections should be small.

Having small α', or string worldsheet, corrections has the invariant meaning of having large curvatures with respect to the string length $\sqrt{\alpha'}$, i.e., $\alpha'/R^2 \ll 1$. For the $AdS_5 \times S^5$ space, that becomes

$$\frac{R^2}{\alpha'} = \sqrt{4\pi g_s N} = \sqrt{g_{YM}^2 N} \gg 1 \,, \tag{23.35}$$

which therefore translates in the SYM theory into a very large 't Hooft coupling,

$$\lambda \equiv g_{YM}^2 N \gg 1. \tag{23.36}$$

Indeed, 't Hooft showed that for an $SU(N)$ gauge theory the effective loop parameter is λ, so perturbation theory corresponds to $\lambda \ll 1$. That means that the perturbative gravity limit is an ultra-nonperturbative limit in SYM. Then the AdS/CFT correspondence is an example of *duality*, that relates strong coupling to weak coupling and vice versa.

The next condition is that quantum string corrections are small, which amounts to g_s small, $g_s \ll 1$.

Putting together the two conditions, we obtain $N \to \infty$, which means that the limit is really the one considered by 't Hooft, $N \to \infty$, $g_{YM} \to 0$, but with $\lambda = g_{YM}^2 N$ fixed but very large ($\gg 1$). This 't Hooft limit corresponds in the string side of the correspondence to the gravity limit of string theory.

In terms of the SYM variables, small string corrections correspond to

$$\frac{\alpha'}{R^2} = \frac{1}{\sqrt{\lambda}} \ll 1\ , \quad 4\pi g_s = \frac{\lambda}{N} \ll 1. \tag{23.37}$$

In principle, we have 3 possibilities for the AdS/CFT correspondence:

- (weakest) Matching between the two sides only in the $\alpha'/R^2 \to 0$, $g_s \to 0$ limit.
- (stronger) α'/R^2 corrections match, but g_s corrections don't.
- (strongest) both α'/R^2 and g_s corrections match, so matching of the two sides is at the full string theory level, or equivalently for any N and g_{YM}.

As we said, the strongest form of the correspondence is actually believed to be correct, and the matching is at the full string level. Many α' and g_s corrections have been calculated and found to match.

23.5 Operator/State Map and GKPW Construction

Having been described what the equivalence is, we now proceed to describe the mapping of objects between the two sides.

The first observation is that the symmetries of the gravitational $AdS_5 \times S^5$ background and of the $\mathcal{N} = 4$ SYM CFT match between the two sides: the $SO(4, 2)$ isometry group of AdS_5 becomes the conformal group in $(3, 1)$ flat Minkowski dimensions, and the $SO(6)$ isometry group of S^5 becomes the $SO(6)$ global "R-symmetry" group of the $\mathcal{N} = 4$ SYM theory.

Then the fields in AdS_5, in the KK expansion on the compact space S^5, thus in representations of the isometry group $SO(6)$, should correspond to *gauge invariant operators* $\mathcal{O}$ in the SYM, in the same representation of the global symmetry group $SO(6)$. More precisely, KK expanding *scalar* fields as

$$\phi(x, y) = \sum_n \sum_{I_n} \phi^{I_n}_{(n)}(x) Y^{I_n}_{(n)}(y) \ , \tag{23.38}$$

where $Y^{I_n}_{(n)}(y)$ are spherical harmonics on S^5 in the representation I_n of $SO(6)$, with level n, measuring the eigenvalue of $\Box$ on S^5, the fields $\phi^{I_n}_{(n)}(x)$ in AdS_5, of mass m, correspond to the operators $\mathcal{O}^{I_n}_n$, with conformal dimension

$$\Delta = \frac{d}{2} + \sqrt{\frac{d^2}{4} + m^2 R^2} \tag{23.39}$$

and d is the dimension of the CFT, for us $d = 4$. For p-form fields in AdS_{d+1}, we have the relation

$$(\Delta - p)(\Delta + p - d) = m^2 R^2 \ , \tag{23.40}$$

solved by

$$\Delta = \frac{d}{2} + \sqrt{\frac{d^2}{4} + m^2 R^2 + p(p-d)}. \tag{23.41}$$

For $m = 0$, we obtain

$$\Delta = \frac{d}{2} + \frac{d - 2p}{2} = d - p. \tag{23.42}$$

One important observation is that in AdS space, m^2 need not be just positive as in flat space, which is a condition from normalizability at infinity of localized excitations: $e^{ik\cdot x} = e^{imt+\cdots} = e^{+t\sqrt{-m^2}+\cdots}$ gives $m^2 \geq 0$. In AdS space, normalizability of localized excitations gives the so-called *Breitenlohner–Freedman bound*, which for scalars is

$$m^2 R^2 \geq -\frac{d^2}{4}. \tag{23.43}$$

Among the possible field-operator pairs, two will be of importance to us:

- A *Noether current* J_μ for a global symmetry group G corresponds to a gauge field A_μ for the same symmetry group in the dual gravitational background, just that now it is *local*. From the above general analysis for $p = 1$ (one-form field), we see that a massless gauge field corresponds to a current J_μ of dimension $d - 1$, as it should be. Gauging a global symmetry involves adding a coupling term $\int d^d x J^\mu(x) A_\mu(x)$, so this is a consistent choice.
- An important particular case of the above is the case of the *energy-momentum tensor* $T_{\mu\nu}$, the Noether current for translations, with generator P_μ. The gauge field for them, making locally translation invariant fields, i.e., general coordinate transformation invariant ones, is known to be the metric $g_{\mu\nu}$. Indeed, making a translationally invariant theory into a general coordinate invariant one involves adding a term $-\frac{1}{2}\int d^d x \sqrt{-g} T_{\mu\nu}(x) \delta g^{\mu\nu}(x)$, justifying the mapping.

Having described the map from SYM operators to gravitational states (fields), we should describe how to relate observables on the two sides. This was done by Gubser, Klebanov, Polyakov, and separately by Witten, thus we call it the GKPW construction. Or, since the map described here is the one of Witten, simply the Witten map.

A massless field in AdS space, $\phi(\vec{x}, x_0)$, can be shown to go to a constant at the boundary $x_0 \to 0$, $\phi_0(\vec{x})$, which means that the natural interpretation of $\phi_0(\vec{x})$ is as a source for the boundary operator $\mathcal{O}$ that couples to it. Indeed, we saw above that on the boundary, gauging a symmetry meant adding a term $\int d^d x J^\mu(x) A_\mu(x)$, where A_μ is the source for the current J^μ, and also the boundary value of the field in the dual gravitational background. For massive fields, and other types of fields (general p-forms, fermions, etc.), the analysis is a bit more complicated, but it is similarly done.

In order to define observables in the $\mathcal{N} = 4$ SYM, we are led to define the partition function for the operator $\mathcal{O}$, with a source ϕ_0 for it, which in Euclidean signature is

$$Z_{\mathcal{O}}[\phi_0] = \int \mathcal{D}[\text{SYM fields}] e^{-S_{\text{SYM}} + \int d^d x \mathcal{O}(x)\phi_0(x)}. \tag{23.44}$$

The correlation functions of $\mathcal{O}$ are found from it by differentiation,

$$\langle \mathcal{O}(x_1) \dots \mathcal{O}(x_n) \rangle = \frac{\delta^n}{\delta\phi_0(x_1) \dots \delta\phi_0(x_n)} Z_{\mathcal{O}}[\phi_0] \Big|_{\phi_0=0}. \tag{23.45}$$

Since AdS/CFT relates the CFT theory with the AdS one, it is natural to propose that the partition function above is the same as the partition function in AdS space, with boundary value $\phi_0(\vec{x})$ for the fields $\phi(\vec{x}, x_0)$, i.e.,

$$Z_{\mathcal{O}}^{\text{CFT}}[\phi_0] = Z_{\text{string}}^{\text{AdS}}[\phi[\phi_0]]. \tag{23.46}$$

This should be valid generally in the string theory case, but in particular in the supergravity limit of string theory (for $\alpha' \to 0$, $g_s \to 0$), the string partition function should become just the exponent of the supergravity action, on-shell for the fields ϕ as a function of their boundary values ϕ_0, thus for $\phi = \phi[\phi_0]$, that is,

$$Z_{\text{string}}^{\text{AdS}}[\phi[\phi_0]] \to e^{-S_{\text{supergravity}}[\phi[\phi_0]]}. \tag{23.47}$$

Finally then, the Witten map in the supergravity limit is

$$Z_{\mathcal{O}}^{\text{CFT}}[\phi_0] = Z_{\text{string}}^{\text{AdS}}[\phi[\phi_0]] = e^{-S_{\text{supergravity}}[\phi[\phi_0]]}. \tag{23.48}$$

23.6 Generalization to Non-conformal Theories: Gauge/Gravity Duality

We will be interested in a somewhat more general set-up than the one of $AdS_5 \times S^5$ versus $\mathcal{N} = 4$ SYM, which have conformal symmetry and maximal supersymmetry, so we must generalize for other cases.

First, we note that there are other cases, for other dimensions, with conformal symmetry and (near) maximal supersymmetry: the "ABJM model", a 2+1 dimensional model with $\mathcal{N} = 6$ supersymmetry and an $SU(N) \times SU(N)$ gauge group versus string theory in $AdS_4 \times \mathbb{CP}^3$; the "(0, 2) theory" (named after the amounts of supersymmetry) in 6 dimensions versus M-theory in $AdS_4 \times S^7$ background. But then it was realized that we can consider more general cases, with either less supersymmetry, or with no conformal symmetry, and still a version of the correspondence is valid, under the new name of the "gauge/gravity duality", the gravitational background being called the "gravity dual".

Some relevant properties of these gravity duals that will be of use are:

(1) The gauge theory lives at the boundary of the gravity dual, usually reduced on a compact space X_d, giving a holographic duality, between a $d + 1$ dimensional gravity theory and a d dimensional gauge theory.

(2) The map is a duality, relating a strongly coupled theory on one side of the duality with a weakly coupled one on the other. In particular, we must have a large N (rank of the gauge group) gauge theory, corresponding to a $g_s \to 0$ theory, and a large $\lambda = g_{YM}^2 N$ theory, corresponding to an $\alpha' \to 0$ limit.

(3) The symmetries must match, so global symmetries of the field theory match the local symmetries in the gravity dual, which usually come, via KK expansion, from the global symmetries of the compact space.

(4) The energy scale in the field theory is $U = r/\alpha'$, where r is the holographic coordinate in the gravity dual. Therefore one geometrizes motion in energy (RG flow) as motion in an extra coordinate.

(5) Supergravity fields in the gravity dual correspond to gauge invariant field theory operators made up from adjoint fields, i.e., "glueball" type operators.

(6) The mass spectrum of a tower of glueballs associated with an operator $\mathcal{O}$ correspond to the mass spectrum of the wave equation for the supergravity field ϕ coupling to $\mathcal{O}$ in the gravity dual.

Important Concepts to Remember

- Anti-de Sitter (AdS) space is a maximally symmetric, homogenous and isotropic space of constant negative curvature, with negative cosmological constant.
- Lobachevsky space is the Euclidean signature space equivalent to the sphere, but embedded in Minkowski space. de Sitter space is the Minkowskian signature equivalent of the sphere, and Anti-de Sitter space is the Minkowskian signature equivalent of Lobachevsky space.
- AdS space has a Poincaré metric, with boundary Poincaré symmetry, that only covers a Poincaré patch, and a global metric, that covers the whole space.

- The boundary of AdS space, in both Poincaré and global metrics, is a finite time away (by a light signal).
- The boundary of Poincaré AdS is flat $\mathbb{R}^4$, and of global AdS is $S^3 \times \mathbb{R}_t$, which are related by a conformal factor in Euclidean signature.
- Conformal transformations are transformations of flat space that change the metric by a position-dependent scale factor, $d\vec{x}'^2 = \Omega^2(x) d\vec{x}^2$. In 2 dimensions, they form an infinite group, but in $d > 2$ Minkowski space, they form the group $SO(d, 2)$.
- The first observation leading to AdS/CFT is the holographic property of AdS space, and the fact that the symmetry group of AdS_{d+1}, $SO(d, 2)$, is the same as the conformal group in d dimensions.
- We derive AdS/CFT heuristically, by considering two points of view for D-branes: as the endpoints of open strings, and as gravitational objects that curve space.
- In the decoupling limit $\alpha' \to 0$ and $r \to 0$, we obtain the equivalence of $\mathcal{N} = 4$ SYM in 3+1 dimensions, with the gauge group $SU(N)$, to string theory in $AdS_5 \times S^5$ background, with N the charge in the background, and $4\pi g_s = g_{YM}^2$.
- The general proposal, found until now to be true, is that the equality is valid at all N and g_{YM}, but initially it was thought that it is true only for small g_s and α', that is, in the classical supergravity limit, corresponding to $N \to \infty$, $g_s \to 0$, with $\lambda = g_{YM}^2 N \gg 1$ and fixed. The correspondence is $\alpha'/R^2 = 1/\sqrt{\lambda}$ and $4\pi g_s = \lambda/N$.
- Gauge invariant operators in $\mathcal{N} = 4$ SYM correspond to fields in AdS (KK dimensionally reduced on S^5), with the same symmetry, and dimension Δ related to the mass squared $m^2 R^2$ of the dual field.
- A Noether current corresponds to a gauge field for the same symmetry group, and the energy-momentum tensor to the gravity field (metric).
- The CFT path integral for operators $\mathcal{O}$ with sources ϕ_0 corresponds to the string path integral with boundary value ϕ_0 for the field ϕ dual to $\mathcal{O}$, $Z_{\mathcal{O}}[\phi_0] = Z^{\text{AdS}}_{\text{string}}[\phi_0]$. In the supergravity limit, we obtain $e^{-S_{\text{sugra}}[\phi[\phi_0]]}$. Correlators are obtained by differentiation with respect to ϕ_0.
- AdS/CFT is generalized to gauge/gravity duality, for dual pairs of a field theory and a gravitational background.

Further reading: The AdS/CFT correspondence was defined by Maldacena in [24]. The prescription for calculating correlation functions was defined by Gubser, Klebanov and Polyakov in [25] and by Witten in [26]. For a review of the AdS/CFT correspondence, see the books [27, 28].

Exercises

(1) Show that the coordinate transformation

$$\begin{aligned} X_0 &= R \cosh\rho \cos\theta \\ X_i &= R \sinh\rho \Omega_i \\ X_{d+1} &= R \cosh\rho \sin\tau \end{aligned} \tag{23.49}$$

is an embedding of the AdS space in global coordinates (23.12) inside the flat space embedding in signature $(d-1,2)$.
(2) Show that the coordinate transformation $x'_\mu = x_\mu + v_\mu$ of flat space, with v_μ in (23.22), is a conformal transformation with $\Omega = 1 - \sigma_v(x)$, where $\sigma_v = \partial_\mu v^\mu / d$.
(3) Solve the KG equation $(\Box - m^2)\phi = 0$ in Poincaré coordinates of AdS_{d+1}, near its boundary at $x_0 = 0$, and show that the two independent solutions are $\sim x_0^{2h_\pm}$, with

$$2h_\pm = \frac{d}{2} \pm \sqrt{\frac{d^2}{4} + m^2 R^2}. \tag{23.50}$$

(4) Consider the supergravity solution for N extremal M5-branes,

$$\begin{aligned} ds^2 &= f_5^{-1/3}(r)[-dt^2 + d\vec{x}_5^2] + f_5^{2/3}(r)[dr^2 + r^2 d\Omega_4^2] \\ f_5(r) &= 1 + \frac{\pi N l_P^3}{r^3} \\ F_4 &= *(dt \wedge dx^1 \wedge \cdots \wedge dx^5 \wedge df_5^{-1}). \end{aligned} \tag{23.51}$$

Show that the decoupling limit $r \to 0, l_P \to 0$, with $U^2 \equiv r/l_P^3$ fixed gives an AdS/CFT duality, with an $AdS_7 \times S^4$ metric and constant F_4. Calculate R_{S^4} and R_{AdS_7}.
(5) Consider the classical solution for a scalar ϕ_i in AdS_{d+1} space in Poincaré coordinates, written as a function of the boundary value ϕ_{0i} as

$$\phi_i(z_0, \vec{z}) = \int d^d z K_{B,\Delta_i}(z_0, \vec{z}; \vec{x}) \phi_{0i}(\vec{x}). \tag{23.52}$$

Use the Witten (GKPW) prescription in the classical supergravity limit, in order to show that, if we have 3 scalars ϕ_1, ϕ_2, ϕ_3 with interaction term $\mathcal{L}_{\text{int}} = \lambda \phi_1 \phi_2 \phi_3$, the 3-point correlator for the operators dual to it is

$$\langle \mathcal{O}_1(x_1)\mathcal{O}(x_2)\mathcal{O}(x_3)\rangle = -\lambda \int \frac{d^d z dz_0}{z_0^{d+1}} K_{B,\Delta_1}(z_0, \vec{z}; \vec{x}_1) K_{B,\Delta_2}(z_0, \vec{z}; \vec{x}_2) K_{B,\Delta_3}(z_0, \vec{z}; \vec{x}_3). \tag{23.53}$$

Part III
Introduction to String Cosmology

Chapter 24
Problems of String Inflation

We now move on to the study of string cosmology, i.e., applications of string theory to cosmology. We have to start, of course, with trying to obtain the most popular cosmological model, inflation. But it turns out that strictly speaking (applying the fully top-down mentality of deriving everything from string theory), we have no good model of string inflation yet. That is to say, there are models that when we consider some approximations seem to give inflation, but in the end, corrections not taken into account can spoil the nice features. Therefore in this chapter we will examine the problems that appear when we try to apply string theory to find inflationary models.

The first question to ask is, why do we want to obtain string inflation? Of course, inflation being the most popular cosmological model we need to see whether it needs to be embedded into something else. It turns out that in fact, inflation is quite UV sensitive. We will review in more detail the relevant argument when discussing the Brandenberger-Vafa "string gas cosmology", but the crux of it is that, for a sufficiently large number of e-folds of inflation, the last scales to re-enter the horizon (the CMBR scales) spent a large part of the early history (before going outside the horizon and being frozen-in) being sub-Planckian, hence they must have necessarily been influenced by trans-Planckian physics, i.e., quantum gravity. While this does not happen for models with only 50–60 e-folds of inflation, it points out the UV sensitivity of inflation, hence the need to have it at the very least embedded into a consistent theory of quantum gravity.

Inflation is an effective field theory description, which means one needs to integrate out all modes of the string higher than a cut-off scale. In phenomenological effective field theory, one writes all the operators consistent with symmetries, valid below a cut-off, and puts arbitrary coefficients for them, to be fixed by experiments, but in a top-down approach like the string theory one, one must integrate out higher modes, and obtain the correct coefficients in the effective field theory.

What do we need to have in order to obtain inflation? Inflation was presented as a more general paradigm, for accelerated expansion governed by evolution in a certain field, but the most popular incarnation is in terms of a scalar field that slowly rolls down a flat potential. Therefore, we need to obtain a scalar with a nearly flat potential.

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_24

But the problem in string theory is not to obtain such a scalar, but rather that we have too many, coming from the various parameters of compactification. These are *moduli*, i.e., scalars that perturbatively have no potential (it takes no energy to deform the compactification along these directions), but could have a nonperturbative potential.

24.1 Moduli of Type IIB String Compactifications on CY_3

In particular, for a popular concrete example, we will consider type IIB theory compactified on a CY_3 space. We will not describe the most general topological description of the Calabi-Yau spaces. We will just say some things that we need.

The CY_3 space has a number of topologically nontrivial 3-dimensional surfaces, or 3-cycles, the number of independent ones being given by b_3, the "third Betti number", characterizing *homology* (the nontrivial way in which surfaces interesect). A homology basis for 3-cycles is given by (A_I, B^J), where $I, J = 1, \ldots, b_3/2$. We can define the *intersection number of surfaces* as the number of nontrivial "encirclings" of one surface by the other. It is defined as an antisymmetric function of surfaces, and then we have (for the so-called A-cycles and B-cycles)

$$A_I \cap B^J = -B^J \cap A_I = \delta_I^J; \quad A_I \cap A_J = 0 = B^I \cap B^J. \tag{24.1}$$

The basis is unique up to symplectic transformations with integer coefficients acting on the cycles (A_I, B^J), i.e., $Sp(b_3, \mathbb{Z})$ transformations, that leave invariant the intersection numbers.

The homology basis above is dual to s basis of 3-forms on the space, for the *cohomology* group $H^3(X, \mathbb{R})$, given by (α_I, β^J), such that

$$\int_{A_I} \beta^J = \delta_I^J; \quad \text{etc.} \tag{24.2}$$

Complex Structure Moduli

Then the complex structure moduli, or shape moduli (in the case of the square 2-torus T^2, the complex structure modulus is R_2/R_1) are obtained as integrals of the unique holomorphic 3-form Ω on the homology basis (A_I, B^J), i.e.,

$$F_I = \int_{A_I} \Omega; \quad Z^J = \int_{B^J} \Omega. \tag{24.3}$$

However, since the normalization of Ω is irrelevant, we actually have $h^{2,1} = b_3 - 1$ independent complex structure moduli (there are some more nontrivial topological numbers called Hodge numbers $h^{p,q}$ that will not be explained here).

Kähler Structure Moduli

The Kähler structure moduli, or size moduli (in the case of the square 2-torus T^2, the Kähler structure modulus is the volume $R_1 R_2$) are obtained in a similar way, but by integrating a unique 2-form on a basis of 2-cycles.

On a Calabi-Yau space there is a unique covariantly constant spinor η, as we saw in the last chapter, and from it we can construct the Kähler form

$$K_{ij} = \bar{\eta}\Gamma_{ij}\eta, \tag{24.4}$$

and in turn the complex structure

$$J^i{}_j = g^{ik}K_{kj}\ , \tag{24.5}$$

or sometimes written as a 2-form (since it is antisymmetric)

$$J = g_{i\bar{j}}dz^i \wedge dz^{\bar{j}}. \tag{24.6}$$

But then in string theory we also the NS-NS 2-form field $B_{\mu\nu}$, so we can define the *complexified Kähler class*

$$K = J + iB. \tag{24.7}$$

We can define also the homology of 2-cycles, by finding the basis of topologically nontrivial 2-dimensional surfaces (2-cycles) $(A'_{I'}, B'^{J'})$, $I', J' = 1, \ldots, b_2/2$, with orthonormal intersection numbers

$$A'_{I'} \cap B'^{J'} = \delta^{J'}_{I'}, \quad A'_{I'} \cap A'_{J'} = 0 = B'^{I'} \cap B'^{J'}. \tag{24.8}$$

Here there are $b_2 = h^{1,1}$ 2-cycles, where b_2 is the second Betti number (and $h^{1,1}$ is a Hodge number).

Then as before, the Kähler structure moduli are given by the integrals of the 2-form K over the basis of 2-cycles,

$$X_{I'} = \int_{A_{I'}} K, \quad X^{J'} = \int_{B^{J'}} K. \tag{24.9}$$

These 2 types of modulis exist for whatever CY_3 construction, but in more general cases we can have other types of moduli as well, like:

- **D-brane moduli**, coming from the fields on the worldvolume of D-branes. The most obvious one is the position X_0 of the D-brane in the transverse space, but in general there could be others as well. For instance, a D-brane can wrap a variable cycle in the geometry, parametrized by some scalar field.
- **axions**, coming from fluxes of antisymmetric tensor fields on cycles in the geometry.
- there can be also other moduli in more complicated cases.

Example

One interesting example is provided by the type IIB flux compactification, with G-flux.

We define the complex scalar made up from the RR axionic scalar a and the dilaton ϕ,

$$\tau = a + ie^{-\phi}. \tag{24.10}$$

The G-flux refers then to the combination of the NS-NS 3-form field strength H^{NS-NS} and the R-R 3-form field strength F^{R-R},

$$G = F^{R-R} - \tau H^{NS-NS}. \tag{24.11}$$

In the case that the G-flux is nontrivial on cycles, we have the GVW superpotential

$$W = \int_{CY_3} \Omega \wedge G \ , \tag{24.12}$$

where Ω is the holomorphic 3-form.

We see that we have very many scalar moduli in a string compactification, in fact *too many*. In fact, we have not observed a single light scalar in the real world, which means that we need to stabilize the string moduli, by a nonperturbative (super) potential.

Since today we don't see any light scalars, it means that today *all* the moduli need to be stabilized. There is one possibility around that will be explored towards the end of the book, in the form of so-called chameleon scalars. They arose out of the observation that in fact, we only know that there are no light scalars here on Earth (where we can perform fifth force search experiments, and see that there is none), or in the Solar System (the motion of the planets follows from Newtonian theory, without the need for a fifth force). But we could still have a scalar (like the chameleon) that is light on cosmological scales and massive on Earth and Solar System scales.

But at early stages, we still want one of the scalars to be not stabilized, namely the inflaton. In fact, in the usual picture of new inflation, we would have a flat plateau, and after a drop, we have a minimum of large mass, stabilizing it now.

However, note that this presents a problem, because we want to have all the moduli stabilized, and moreover stabilized at zero potential (no cosmological constant), whereas without the stabilization, the extra moduli would have no potential, that means that we should extend the inflaton potential trivially in all extra moduli directions. But that means that generically the quantum corrections must introduce steep drops (followed by stabilization) in transverse directions, invalidating the flatness of the potential needed for inflation, since the evolution in field space is along the steepest path.

To check whether the above happens, or the potential continues to have flat valleys, we would need to have the full nonperturbative moduli potential (since stabilization is nonperturbative), which would in effect mean that we have to solve the full nonperturbative string theory in some sector, which is clearly impossible. That is why

we don't have a good model of string inflation yet: to stabilize moduli we need non-perturbative corrections, so g_s is large(ish), but then we must consider *all* possible corrections, since without it we are not sure whether we are in a true valley, or on top of a hill with a flat narrow top.

24.2 Approximations for String Calculations of the Potential

In doing actual calculations for the potential, one usually does a combination of approximations. The most important approximations, well established in the literature over many years, are as follows.

(1) The α' *expansion,* needed in order to have a perturbative string worldsheet. In this context, it is an expansion in

$$\alpha'/\mathcal{R}^2 \ll 1 \ , \tag{24.13}$$

where $\mathcal{R}$ is the scale of variation of the various background fields. It is usually hard to satisfy.

(2) The *string loop expansion,* in $g_s \ll 1$, in order not to have the full nonpeturbative quantum string theory. However, as said above, we need to at least partially contradict it for moduli stabilization. This presents a problem, since we must invoke some nonperturbative contributions that stabilize the moduli, yet we cannot calculate the full nonperturbative potential.

(3) The *probe approximation* for D-branes. When treating D-branes, usually one considers them as probes, i.e., they do not backreact on the background. They are only fixed at some point, or wrapping some cycles, in the compact manifold. But often times, the approximation is, or must be, violated.

(4) An opposite approximation is also possible, considering a *large charge approximation,* $N \gg 1$, for the number of branes, which means that the branes drastically curve the background, and one usually ends up in a near-horizon geometry, like in the case of large N number of D3-branes, when we obtain the $AdS_5 \times S^5$ geometry.

(5) The *smeared approximation.* One other possibility to deal with branes, in the case that it is difficult to find the dependence on one of the coordinates, x, is to consider the brane charge to be uniformly distributed, or *smeared* in that direction, such that the fields are independent of x.

(6) The *adiabatic approximation* amounts to a slow motion of branes, done such as not to change the background.

(7) More of a subset of number 6 is the *moduli space approximation*, a general procedure in the case of solitonic objects. When a static configuration of solitons (or here, D-branes) is stable (has no potential), it means that the various positions are moduli (this can be generalized to other kinds of moduli). But if we give a small (relative) velocity to the moduli (i.e., make them time dependent), we don't

have anymore a solution. Instead, one must re-solve the equations of motion and introduce first order corrections in the velocities (time dependence), and one can find an effective Lagrangian on the moduli space, of the type

$$L = g_{ij}\dot{X}^i\dot{X}^j - V(X^k) \ , \tag{24.14}$$

where X^i are the moduli and g_{ij} is the moduli space metric.

(8) *Mode truncation* (for instance, truncation of KK modes).

What usually happens is that at least one of the above approximations contradicts the needed regime of parameters, so we cannot trust the result.

24.3 Eta Problem in String Theory

The problem is that the stabilization of moduli doesn't decouple from inflation. One considers a model of "string inflation" that works, in the absence of moduli stabilization, but after stabilizing the moduli, we create terms that spoil the flatness of the potential. The "eta problem" is the fact that the corrections introduced this way, or perhaps by quantum corrections (see later) spoil the smallness of the inflationary parameter η (which needs to be $\ll 1$ for slow roll).

Example

Consider open strings stretched between two sets of branes, D-brane sector A and sector B. As we already said, to obtain the effective field theory of inflation, we must integrate out the string modes above a cut-off. Integrating out the open string, of mass M_{AB}, we obtain terms of the type

$$\frac{\mathcal{O}_A^{(\Delta_1)}\mathcal{O}_B^{(\Delta_2)}}{M_{AB}^{\Delta_1+\Delta_2-4}} \ , \tag{24.15}$$

such as to give a dimension 4 operator. Here $\mathcal{O}_A$ is an operator for fields in the D-brane sector A, of dimension Δ_1 and $\mathcal{O}_B$ in D-brane sector B, of dimension Δ_2.

For instance, consider $\mathcal{O}_A$ to be a constant (independent of the inflaton, rather) V_0, and $\mathcal{O}_B$ to be a mass term for the inflaton, ϕ^2. Since $M_{AB} = \Delta X/(2\pi\alpha')$, and if $\Delta X < \sqrt{\alpha'}$ we need string corrections, we must instead consider $\Delta x > \sqrt{\alpha'}$, which means that $M_{AB} \sim M_{\rm Pl}$ is natural, leading to a term in the effective field theory of type

$$\sim \mathcal{O}_4\frac{\phi^2}{M_{\rm Pl}^2}. \tag{24.16}$$

Then then η parameter of inflation is

$$\eta \equiv M_{\rm Pl}^2\frac{V''}{V} \sim \frac{M_{\rm Pl}^2}{\phi^2}. \tag{24.17}$$

This is generically > 1, but moreover, usually one has also independently $\mathcal{O}_4$, so that in total

$$\mathcal{O}_4 \left(1 + \frac{\phi^2}{M_{\rm Pl}^2}\right) , \tag{24.18}$$

which gives $\eta = 1$, or more generally $\eta \sim \mathcal{O}(1)$.

This is the eta problem, at the classical level (from the point of view of the effective field theory). However, there are also (perhaps even more importantly) quantum corrections, that will be analyzed later.

24.4 Energy Scales for Most Controlled Approximations

In practice, we need to have some sort of control over the approximations made in order to calculate (so that we can trust them). In that case, several energy scales appear, in a particular order, that I will explain in the following.

- In general, we have the Planck scale larger than the string scale, $M_{\rm Pl} > M_s = 1/\sqrt{\alpha'}$. This is generic for *perturbative* string theory, i.e., with $g_s \ll 1$, which as we said is needed for the absence of large quantum string corrections. The 4 dimensional Planck scale appears from the coefficient of the Einstein action in 4 dimensions, $M_{\rm Pl}^2/2$. But that arises from KK dimensional reduction from 10 dimensions of the same term, with coefficient $1/(g_s^2 2\kappa_{N,10}^2)$. The $1/g_s^2$ is there since this is a closed string action, and so must have this overall coefficient (g_s is the closed string coupling), and the $1/2\kappa_{N,10}^2$ is the standard gravity coupling, which by dimensional analysis must equal $M_s^8/2$. KK dimensional reduction means that we get the volume of the 6 dimensional compact space $V_{\rm compact}^{(6)} = M_{\rm KK}^{-6}$ as a prefactor. Here we have defined a KK scale (scale of masses of KK modes) by analogy with the one-dimensional (circle) case, where $R = 1/M_{\rm KK}$. Finally, that gives

$$M_{\rm Pl}^2 = \frac{1}{\kappa_{N,10}^2} \frac{1}{g_s^2} V_{\rm compact}^{(6)} \Rightarrow M_{\rm Pl} = \frac{M_s}{g_s} \left(\frac{M_s}{M_{KK}}\right)^3 . \tag{24.19}$$

  We see that $g_s \ll 1$ and $M_s \sim M_{\rm KK}$ indeed implies $M_s \ll M_{\rm Pl}$. In a natural compactification, indeed $M_{\rm KK} \sim M_s$ (since the compact space would become compact due to string effects). But if one tries to obtain large extra dimensions, we would have $R \gg \sqrt{\alpha'}$, or $M_s \gg M_{\rm KK}$, which would mean that M_s is even smaller than $M_{\rm Pl}$. The only issue is whether we can have $M_{\rm KK} > M_s$, but that is both unlikely in a general quantum gravity scenario, and also in particular in string theory, we have T-duality equating large and small volume physics, so in that case we can take it as excluding an equivalent scenario.
- Therefore we have also $M_s > M_{\rm KK}$, or $V_{\rm compact}^{(6)} \gg M_s^{-6} = \alpha'^3$. Note that this large volume in string units will also imply small α' corrections, so it is actually a needed condition in order to avoid string worldsheet corrections.

- The energy scale of inflation is the Hubble scale H_i during inflation. For the validity of inflation as an effective field theory, we must have the cut-off scale be larger. In this case, the cut-off refers to $M_{\rm KK}$, beyond which we must consider the whole infinite tower of KK modes, or equivalently the full 10 dimensional theory. Therefore we must have $H_i < M_{\rm KK}$.
- Finally, we usually need to use supersymmetry during inflation, if we are to trust quantum corrections calculated (without susy, there is not much control over quantum corrections). But that means that we must have the scale of susy breaking in 4 dimensions (of the breaking of the $\mathcal{N} = 1$ susy to MSSM) be less than the scale of inflation, $M_{\text{susy br.}} < H_i$.

All in all, we have the needed hierarchy

$$M_{\text{susy br.}} < H_i < M_{\rm KK} < M_s < M_{\rm Pl}. \tag{24.20}$$

Quantum Eta Problem Redux

We now can explain the *quantum eta problem* better, as something akin to the hierarchy problem in the Standard Model, namely that the mass of some scalar (in the case of the Standard Model, the Higgs, in the case of inflation, the inflaton) is naively taken by quantum corrections in a generic effective field theory to the scale of the UV cut-off, $\Lambda_{\text{cut}-\text{off}}$. For that *not* to happen, we need something drastic like supersymmetry that prevents various quantum corrections.

But the effective field theory of inflation is usually only valid in a very small window, since $H_i \in (M_{\text{susy br.}}, M_{\rm KK})$, and as we saw, at $M_{\rm KK}$ we break most of the susy, being left with the $\mathcal{N} = 1$ needed for phenomenology. Yet nevertheless, one obtains usually susy breaking of the remaining $\mathcal{N} = 1$ at a scale smaller, but not much smaller than $M_{\rm KK}$. That means that we have

$$\Lambda_{\text{cut}-\text{off}} \gtrsim H_i \ , \tag{24.21}$$

which in turn means that the quantum corrections for the inflaton mass (loops of other fields inserted into the inflaton propagator) give a mass at this cut-off scale, i.e.

$$\Delta m_\phi \sim H. \tag{24.22}$$

Since moreover the constant part of the potential $\mathcal{O}_4 = V_0$ defines the Hubble scale during inflation by $V_0 = 3M_{\rm Pl}^2 H_i^2$, we have the inflaton potential

$$V \sim \mathcal{O}_4 + \Delta m_\phi^2 \frac{\phi^2}{2} = 3M_{\rm Pl}^2 H_i^2 + \Delta m_\phi^2 \frac{\phi^2}{2} \ , \tag{24.23}$$

leading to

$$\Delta\eta = \frac{\Delta m_\phi^2}{3H^2} \sim 1. \tag{24.24}$$

Again, to see whether or not this happens at the quantum level (in the effective field theory of inflation, itself obtain by integrating out massive modes beyond $\Lambda_{\text{cut}-\text{off}}$), we would need to have the *full* quantum effective potential. But getting that would be equivalent to solve the full nonperturbative string theory, an impossible task. We would have to be very lucky to have a model where we could calculate everything in a controlled approximation.

24.5 Large Field Inflation

Another issue that can appear in string theory models for inflation is a generic one to most approaches, related to the Lyth bound, found in Chap. 11 (of part I). We saw then that, for a tensor to scalar ratio of perturbations r and a number of e-folds of inflation N_e, we have a minimum inflaton field excursion

$$\Delta\phi_{\Delta N_e} \geq M_{\text{Pl}}\sqrt{r}\frac{\Delta N_e}{\sqrt{8}}. \tag{24.25}$$

This would become a problem if we see a large(ish) r (as of this moment, we have not observed tensor modes, only scalar modes), since it would imply that we need to have a $\Delta\phi$ that is trans-Planckian.

The first problem with that is that it might not be kinematically allowed. In fact, a typical string compactification (with string scale extra dimensions) usually doesn't allow it.

In something like large extra dimensions this would be allowed. In fact, large extra dimensions means that the scalar representing the volume in string units *is* (by definition) trans-Planckian. But some models are extremely constrained (see for instance my papers with Khoury and Hinterbichler, and Khoury, Hinterbichler and Rosenfeld), to the point that it is not clear that there is a good model compatible with experiments.

But moreover, even if one has a kinematically allowed trans-Planckian $\Delta\phi$, that would mean that we need to control dynamically the flatness of the inflaton over these large excursions. But in a generic effective field theory, we would expect corrections to the operator of dimension four (the potential) $\mathcal{O}_4$, of the order of

$$\mathcal{O}_4\left(\frac{\phi}{M_{\text{Pl}}}\right)^n, \tag{24.26}$$

which would naively blow up. So unless they are suppressed by unknown quantum gravity (string theory) mechanisms, these would make any calculation of the potential unreliable.

One set of models that has become popular recently (to be considered later on in the book) is called *axion monodromy* inflation, and is the only known consistent string model with trans-Planckian field excursion where it seems like the quantum corrections are under control.

Important Concepts to Remember

- For string inflation one of the scalar moduli of string theory is the inflaton.
- In string theory, we have complex structure moduli, Kähler moduli, D-brane moduli, axions, etc. Any of them could be the inflaton, with the other ones stabilized.
- But that means that we need an almost flat direction, that cannot have a steep drop in any transverse direction, which is what would generically happen when we stabilize the transverse directions.
- Among approximations that one must use in string theory in order to be able to calculate the potential, we have the α' and g_s string expansions, either the probe or large charge approximation for branes, the smeared approximation, the adiabatic approximation, the moduli space approximation, and KK mode truncations.
- The eta problem in string theory is that quantum corrections to η are of order one, since generic quantum corrections to the potential are of order one in Planck units. They also appear when we try to stabilize the moduli (which is, necessarily, nonperturbative).
- In order to control the appoximations, we need the hierarchy of scales $M_{\rm susybr.} < H_i < M_{\rm KK} < M_s < M_{\rm Pl}$.
- The cut-off scale for quantum corrections must obey $\Lambda_{\rm cut-off} \gtrsim H_i$, leading to $\Delta m_\phi \sim H$, so $\delta\eta \sim 1$.
- The Lyth bound says that if the tensor to scalar ratio of fluctuations r is found to be large, and ΔN_e is also large, we will have a large field excursion $\Delta\phi$.
- That is hard to obtain in string models, and also can generate large quantum corrections, leading to the η problem, so we would need to explain why that doesn't happen.

Further reading: See the string inflation reviews [29, 30], as well as Sects. 4.1, 4.2 in [31] (arXiv review containing part of the book [32]).

Exercises

(1) Consider an $\mathcal{N} = 1$ susy model with chiral superfield Φ, canonical Kähler potential $K = \Phi^\dagger\Phi$ and superpotential $W = A\Phi^3$ coupled to supergravity, with $|A| \ll 1$. Does it have a *classical* η problem?
(2) Assuming that we see an r (tensor to scalar ratio of perturbations) at the current limit, $r \simeq 0.10$, and we have 60 e-folds of inflation, what is the reduction needed from *generic* quantum corrections to operators $\mathcal{O}_4$?
(3) Consider compactification on the CY_3 space $(T_2)^{\otimes 3}$ (a product of 3 tori), and no D-branes or fluxes. How many moduli there are?
(4) Consider an axion, which is a scalar field with a potential of the type

$$V(\phi) = V_0[1 - \cos(\phi/f)]. \tag{24.27}$$

Does it have a classical eta problem? If we see a tensor to scalar perturbation ratio r, what is the bound on the constant f such that we don't have transplanckian (large field) quantum corrections?

Chapter 25
Problems of the Supergravity Approximation to String Inflation

Because string theory itself is not very calculable, there are few truly stringy models of cosmology. Therefore, one mostly considers the low energy limit of string theory, supergravity, as an effective field theory, and in this chapter we want to note the problems we encounter when we try to obtain inflation in this effective field theory. As usual, for phenomenological reasons we only consider $\mathcal{N} = 1$ supersymmetry, considering that the rest has been lost at a high scale. In $\mathcal{N} = 1$ supersymmetry, besides the supergravity multiplet, we have the chiral multiplet Φ and the vector multiplet V^a.

25.1 Supergravity Plus Chiral Multiplet

We start by the most standard case, of $\mathcal{N} = 1$ supergravity coupled to a scalar (chiral) multiplet Φ. As we saw in part II, the potential for supergravity coupled to chiral superfields Φ^i is

$$V = e^{K/M_{\rm Pl}^2}\left[g^{i\bar{j}}D_i W\overline{D_j W} - \frac{3}{M_{\rm Pl}^2}|W|^2\right]. \tag{25.1}$$

Here $K(\Phi^i, \bar{\Phi}^i)$ is the Kähler potential, giving the kinetic term, and $W(\Phi^i)$ is the superpotential, which is a holomorphic function of its arguments, and the covariant derivative is

$$D_i W = \partial_i W + \frac{\partial_i K}{M_{\rm Pl}^2}W. \tag{25.2}$$

In the following, we will only consider a single chiral superfield Φ.

The simplest assumption is to consider the Kähler potential to be Taylor expandable around zero, and write

$$K(\Phi, \bar{\Phi}) = K_0 + \partial_\Phi \partial_{\bar{\Phi}} K|_0 \Phi\bar{\Phi} + \cdots \tag{25.3}$$

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_25

Note that this expansion is not the most general one. We have ignored the possible linear term $\Phi + \bar{\Phi}$ in the expansion, on the grounds that it gives zero metric on field space, but in principle it can be there, and we will see at the end that an important special case has such a nonzero term.

With the above Kähler potential, we obtain the scalar kinetic term

$$-(\partial_\phi \partial_{\bar{\phi}} K)\partial_\mu \phi \partial^\mu \bar{\phi} \,, \tag{25.4}$$

and the scalar potential has the terms

$$V = V_0 \left(1 + g_{\phi\bar{\phi}} \frac{\phi\bar{\phi}}{M_{\rm Pl}^2} + \cdots\right) + \cdots \,, \tag{25.5}$$

coming from the exponential prefactor $e^{K/M_{\rm Pl}^2}$ independently of the superpotential W. But these terms by themselves will give a $\Delta\eta = g_{\phi\bar{\phi}} \sim \mathcal{O}(1)$, i.e., a *superpotential independent* order one contribution to η, as we can easily see. Here

$$V_0 = e^{K_0} \left[(\partial_\phi \partial_{\bar{\phi}} K)^{-1} D_\phi W \overline{D_\phi W} - \frac{3}{M_{\rm Pl}^2}|W|^2\right]\Big|_{\phi=0} \,. \tag{25.6}$$

But moreover, in the low energy limit of string theory (supergravity limit), it is natural to assume that we get a superpotential that, if it can be Taylor expanded, gives

$$W = M_{\rm Pl}^3 \sum_p c_p \left(\frac{\phi}{M_{\rm Pl}}\right)^p \,, \tag{25.7}$$

with $c_p \sim \mathcal{O}(1)$. That in turn, will generically give also contributions of order one to η, at least if we have, together with the first nonzero coefficient, c_p, also the coefficient c_{p+2} to be nonzero (if c_{p+2} is zero, we get a field-dependent η, which can be made to be small; if not, we get a constant that is what it is). Of course, these will interfere with each other in the potential V and with the above generic contribution coming from K, so it could be that they cancel the total contribution to η, but this would then involve some *fine tuning*.

This problem could perhaps be circumvented if there is no simple Taylor expansion for K. For instance, for the volume modulus ρ of the CY_3, a complex Kähler modulus combining the volume and another scalar, the Kähler potential is

$$K(\rho, \bar{\rho}) = -3M_{\rm Pl}^2 \ln[\rho + \bar{\rho}]. \tag{25.8}$$

Then indeed, there is no Taylor expansion around zero. Moreover, if we introduce D3-branes in the CY_3, at positions z_α, $\alpha = 1, 2, 3$ in CY_3, we obtain a modification inside the log,

$$K = -3\ln[\rho + \bar{\rho} - \Delta k(z_\alpha, \bar{z}_\alpha)] \,, \tag{25.9}$$

and we will see later that in a popular model (the KKLMMT model), the position of a D3-brane in the compact space acts as the inflation.

However, as we have said previous chapter, we actually need to take care to stabilize the remaining moduli, so ρ must in particular be stabilized, by the inclusion of a nonperturbative superpotential. That means that ρ gains a large mass around its minimum. But then from K, this will imply a mass for z_α as well.

This is an example of the eta problem in string theory: we have a good inflationary model, but when we remember to stabilize the non-inflaton moduli, we generate too large a mass for the inflaton, destroying inflation.

One could argue that we need some softly broken symmetry to guarantee the near-flatness of the potential. The obvious one is a shift symmetry,

$$\phi \to \phi + \text{const.} \,, \tag{25.10}$$

softly broken by small power law terms in the potential, i.e., with Lagrangian

$$\mathcal{L} = -\frac{1}{2}(\partial_\mu \phi)^2 - \lambda_p \frac{\phi^p}{M_{\text{Pl}}^{p-4}} \,, \tag{25.11}$$

with $\lambda_p \ll 1$. Then at $\lambda_p = 0$ we have an exact shift symmetry, but the potential breaks it softly. However, note that this is really parametrizing what we want to obtain, not really "explaining" anything, so one needs a deeper reason behind the smallness of the coefficients.

25.2 D-Term Inflation

We now move on to the other possibility, of having also an abelian vector superfield V^a, with auxiliary field D^a. Here a is an index in the adjoint representation of the gauge group. The scalar potential in this case is given by solving for the auxiliary field, and replacing it in the "D-term" potential, which for a standard YM kinetic term $-(1/4g^2)F^a_{\mu\nu}F^{a\mu\nu}$ is

$$V_D = \frac{g^2 D^a D^a}{2} \,, \tag{25.12}$$

that is to be added to the F-term considered previously (in rigid susy, the potential for the chiral superfield is given by $F^i F^i/2$, with F^i the auxiliary field in the chiral multiplet).

If we have a coupling of the vector and chiral multiplets Φ^i, with a canonical kinetic term for Φ^i, by solving the equation of motion of the auxiliary field D^a we get

$$D^a = \Phi^\dagger T^a \Phi \equiv \Phi^{\dagger i} (T^a)_i{}^j \Phi_j. \tag{25.13}$$

More generally, for a general gauge kinetic term $-(1/4g_a^2) f_a(\phi) F^a_{\mu\nu} F^{a\mu\nu}$, we obtain the D-term in the potential

$$V_D = \sum_a [\mathrm{Re}\, f_a(\phi)]^{-1} \frac{g_a^2 D_a^2}{2} \,, \tag{25.14}$$

and the auxiliary field solution is

$$D^a = \Phi_i (T_a)^i{}_j \frac{\partial K}{\partial \Phi_j}. \tag{25.15}$$

We can also add to the Lagrangian the *Fayet-Iliopoulos (FI) term* defined by a FI parameter ξ^a,

$$\int d^2\theta d^2\bar{\theta} \xi^a V^a \,, \tag{25.16}$$

which effectively replaces

$$D^a \to D^a + \xi^a \tag{25.17}$$

in the D-term. The presence of the FI term means that we can have a spontaneous breaking of susy, by having a nonzero ground state energy (in a susy theory the energy of the -supersymmetric- vacuum must be zero, since the susy algebra contains $\{Q, Q\} \sim H + \cdots$, and so $Q|\psi\rangle = 0$ implies $H|\psi\rangle = 0$), or at least a nonzero local extremum of the potential.

Example: Fayet-Iliopoulos Model

The original FI model provides a simple example. Consider an Abelian vector multiplet, coupled with two chiral multiplets of $U(1)$ charges ± 1, and with superpotential

$$W = m \Phi_+ \Phi_- . \tag{25.18}$$

For standard kinetic terms, we obtain the potential

$$V = m^2 |\phi_+|^2 + m^2 |\phi_-|^2 + (\xi + e^2 |\phi_+|^2 - e^2 |\phi_-|^2)^2 . \tag{25.19}$$

From its equations of motion, we find that the only extremum is at $\phi_+ = \phi_- = 0$, which gives $V_0 = \xi^2 \neq 0$. That means that we have spontaneous susy breaking, but since the VEVs of the scalars vanish, we don't have a Higgs mechanism, and the $U(1)$ is unbroken.

One can in fact generalize this mechanism to inflation, and consider the $\mathcal{N} = 1$ supergravity multiplet coupled to the vector multiplet and 3 chiral multiplets, a neutral one S, and the two charges ones $\Phi_\pm$, with superpotential

$$W = S \Phi_+ \Phi_- \,, \tag{25.20}$$

and Kähler potential

$$K = |S|^2 + |\Phi_+|^2 + |\Phi_-|^2. \tag{25.21}$$

However one finds that one cannot fix all the constraints on the model. It is left as an exercise to find the potential and its equations of motion.

25.3 Field Redefinitions

We saw that we have only a restricted set of possibilities that give inflation. We willl restrict that further, but here we note that in fact we also have some equivalences. In fact, there is a transformation of the Kähler potential, called Kähler transformations, that preserves the metric on the field space $g_{\phi\bar{\phi}} = \partial_\phi \partial_{\bar{\phi}} K$,

$$K \to K + f_1(\phi) + f_2(\bar{\phi}) \,, \tag{25.22}$$

where f_1 and f_2 are arbitrary functions. If we want to maintain K real, we need to have $f_1 = f_2 \equiv f$, but it is not really needed, we only need observables to be real. In rigid susy, this transformation of K is a symmetry of the scalar potential, however in rigid symmetry we have the $e^{K/M_{\rm Pl}^2}$ prefactor. We can in fact transform also the superpotential, in the real case $f_1 = f_2 = f$, by

$$W \to e^{-f/M_{\rm Pl}^2} W \,, \tag{25.23}$$

such that the scalar potential is still invariant. These redefinitions, being actions on the supersymmetric objects K and W, preserve the supersymmetry of the theory.

We can also make redefinitions of just the scalar fields in the multilplets, for instance by going to a canonical kinetic term for the scalars,

$$-\, g_{i\bar{j}} \partial_\mu \phi^i \partial^\mu \bar{\phi}^{\bar{j}} \equiv -\partial_\mu \tilde{\phi}^i \partial^\mu \tilde{\bar{\phi}}^j \,, \tag{25.24}$$

but this transformation doesn't extend at the level of K and W, so this is a non-supersymmetric transformation (it is actually not a transformation of the superfield Φ either).

25.4 Example with $\epsilon, \eta \ll 1$

Until now we have shown various things that don't work, but we should also show something that does work, and one can have $\epsilon, \eta \ll 1$. Consider a canonical Kähler potential

$$K = \bar{\Phi}\Phi \tag{25.25}$$

and the simplest nontrivial superpotential, a linear term

$$W = A\Phi. \tag{25.26}$$

Then the scalar field metric is trivial, $g_{\phi\bar{\phi}} = 1$, and we obtain

$$DW = a\left(1 + \frac{\bar{\Phi}\Phi}{M_{\rm Pl}^2}\right). \tag{25.27}$$

The scalar potential is therefore

$$\begin{aligned} V &= e^{\frac{\bar{\phi}\phi}{M_{\rm Pl}^2}}\left[|a|^2\left|1 + \frac{\bar{\phi}\phi}{M_{\rm Pl}^2}\right|^2 - \frac{3|a|^2}{M_{\rm Pl}^2}|\phi|^2\right] \\ &= |a|^2 e^{\frac{\bar{\phi}\phi}{M_{\rm Pl}^2}}\left[1 - \frac{|\phi|^2}{M_{\rm Pl}^2} + \frac{|\phi|^4}{M_{\rm Pl}^2}\right]. \end{aligned} \tag{25.28}$$

Considering small fields, $|\phi| \ll M_{\rm Pl}$, the mass term for ϕ cancels between the bracket and the prefactor coming from K, and we obtain

$$V \simeq |a|^2\left[1 + \frac{1}{2}\frac{|\phi|^4}{M_{\rm Pl}^4} + \mathcal{O}\left(\frac{|\phi|^6}{M_{\rm Pl}^6}\right)\right]. \tag{25.29}$$

Considering the inflaton to be the real part of ϕ, $\sigma \equiv {\rm Re}\phi$, the inflationary parameters are now field dependent,

$$\begin{aligned} \epsilon &\simeq \left(\frac{\sigma}{M_{\rm Pl}}\right)^6 \\ \eta &\simeq \frac{6\sigma^2}{M_{\rm Pl}^2}\ , \end{aligned} \tag{25.30}$$

so can be put to be $\ll 1$ by having $\sigma \ll M_{\rm Pl}$.

However, note that, despite having small ϵ and η, we don't have new inflation, since we have no *end of inflation*. Indeed, we see that the potential contains a plateau, as needed for new inflation, but it *increases* away from zero, instead of decreasing (so that inflation can end, when the decrease becomes drastic).

Modified Example

We can in fact modify this example to a correct one, by adding a subleading term to the superpotential, to have

$$W = a\Phi\left(1 - b\Phi^n\right)\ , \tag{25.31}$$

where $n \geq 3$, and keeping the canonical K. We obtain now

$$DW = a\left[1 - (n+1)b\Phi^n + \frac{\bar{\Phi}\Phi}{M_{\rm Pl}^2}(1 - b\Phi^n)\right], \tag{25.32}$$

leading to a scalar potential

$$V = e^{\frac{\bar{\phi}\phi}{M_{\rm Pl}^2}}|a|^2\left[\left|1 - (n+1)b\Phi^n + \frac{\bar{\Phi}\Phi}{M_{\rm Pl}^2}(1 - b\Phi^n)\right|^2 - \frac{3|\phi|^2}{M_{\rm Pl}^2}|1 - b\phi^n|^2\right]. \tag{25.33}$$

Again considering only small fields, we now get

$$V \simeq |a|^2 e^{\frac{|\phi|^2}{M_{\rm Pl}^2}}\left[1 - \frac{|\phi|^2}{M_{\rm Pl}^2} - (n+1)b(\phi^n + \bar{\phi}^n) + \frac{|\phi|^4}{M_{\rm Pl}^4} + \mathcal{O}\left(\frac{\phi^5}{M_{\rm Pl}^2}\right)\right]. \tag{25.34}$$

If $n = 3$ and the inflaton is $\sigma =$Re ϕ , we get

$$V \simeq |a|^2\left[1 - 8b\frac{\sigma^3}{M_{\rm Pl}^3} + \mathcal{O}\left(\frac{\sigma^4}{M_{\rm Pl}^4}\right)\right], \tag{25.35}$$

and if $n = 4$ we get

$$V \simeq |a|^2\left[1 - \left(10b - \frac{1}{2}\right)\frac{\sigma^4}{M_{\rm Pl}^4} + \mathcal{O}\left(\frac{\sigma^5}{M_{\rm Pl}^5}\right)\right]. \tag{25.36}$$

In both cases, we see that we still have no mass term, so we can get $\epsilon, \eta \ll 1$, but now we also have the leading term in the potential going down, not up. That means that we can arrange to have new inflation, with a plateau followed by a drop to a minimum. This model was introduced by Yzawa and Yanagida.

As we have shown in [33], these are the only possibilities for canonical Kähler potential and small field inflation.

Until now we have assumed the existence of a Taylor or Laurent expansion for K and W, but this is not the only possibility. We can also have a singularity at $\phi = 0$ that is not of the form of a finite order pole, but rather is an essential singularity.

In general, such a case is hard to obtain in string theory, but the exception is the case of a log Kähler potential, discussed before,

$$K = -3\ln[\Phi + \bar{\Phi}], \tag{25.37}$$

when the inflaton is the real part of ϕ, $\sigma =$ Re ϕ. This is obtained as follows. In the case of a general KK compactification, the compact space metric must be redefined by a volume field,

$$\bar{g}_{mn}(x, y) = e^{2u(x)}g_{mn}(y), \tag{25.38}$$

and this field $u(x)$ can be written in the general form

$$\mathrm{Re}\phi = e^{cu(x)} \,, \tag{25.39}$$

with c a number. Then one obtains the kinetic term

$$+ \frac{\alpha}{(\phi + \bar{\phi})^2} \partial_\mu \phi \partial^\mu \bar{\phi} \,, \tag{25.40}$$

which is obtained from the Kähler potential

$$K = -\alpha \ln[\Phi + \bar{\Phi}]. \tag{25.41}$$

For the case of a CY_3, $\alpha = 3$.

However, in this case, we find that we cannot obtain inflation, unless we modify K by adding terms inside the log, like in the D3-brane moduli case.

25.5 Special Embedding of Inflationary Potentials in Supergravity

However, it turns out that there is a special embedding of (almost) any inflationary potential inside $\mathcal{N} = 1$ supergravity plus a chiral superfield, at the expense of an unusual Kähler potential,

$$K = -3M_{\mathrm{Pl}}^2 \ln \left(1 + \frac{\phi + \bar{\phi}}{\sqrt{3} M_{\mathrm{Pl}}}\right). \tag{25.42}$$

Note that by expanding in a Taylor series around zero, we obtain

$$K \simeq -\sqrt{3} M_{\mathrm{Pl}}(\Phi + \bar{\Phi}) + \frac{1}{2}(\Phi + \bar{\Phi})^2 + \cdots \simeq \bar{\Phi}\Phi + \left[\frac{1}{2}\Phi^2 - \sqrt{3} M_{\mathrm{Pl}} \Phi\right] + \left[\frac{1}{2}\bar{\Phi}^2 - \sqrt{3} M_{\mathrm{Pl}} \bar{\Phi}\right] + \cdots . \tag{25.43}$$

Then by a Kähler transformation, at this order we can remove the extra terms, and be left with only the canonical potential. But of course, the point is that we do have the linear term (not considered before) *before* the Kähler transformation, and the Taylor series continues ad infinitum, which was essential to the fact that this model was excluded from the previous analysis of a Taylor expansion.

Despite what one might think, the choice of the inflaton is not the real part of ϕ, appearing in the Kähler potential, but rather the imaginary part. The canonically normalized inflaton is

$$\phi_{\mathrm{can}} = \sqrt{2}\, \mathrm{Im}\phi. \tag{25.44}$$

Note then that if we would substitute from the beginning in K having just $\phi_{\rm can}$ (just the imaginary part), we would get $K = 0$. Of course, the correct thing to do is to first take the derivatives, and then put the real part to zero. Thus, the kinetic term for ϕ is

$$
\begin{aligned}
-\partial_\phi \partial_{\bar\phi} K \partial_\mu \phi \partial^\mu \bar\phi &= -\frac{\partial_\mu \phi \partial^\mu \bar\phi}{\left(1+\frac{\phi+\bar\phi}{\sqrt{3}}\right)^2} \\
&= -\frac{1}{2}\partial_\mu \phi_{\rm can} \partial^\mu \phi_{\rm can} \ , \qquad (25.45)
\end{aligned}
$$

where in the last equality, we have put the real part of ϕ to zero.

Then we find

$$
DW = \partial_\phi W + \frac{\sqrt{2}}{1+\frac{\phi+\bar\phi}{\sqrt{3}}}\frac{W}{M_{\rm Pl}} \ , \qquad (25.46)
$$

and if we put the real part of ϕ to zero, we find

$$
DW|_{\phi_{\rm can}} = W'\left(\sqrt{2}i{\rm Im}\phi\right) + \frac{\sqrt{3}}{M_{\rm Pl}} W\left(\sqrt{2}i{\rm Im}\phi\right) . \qquad (25.47)
$$

Note that, *if W is a real function* (with real coefficients), the first term is imaginary, while the second is real, so in the scalar potential the $|W|^2$ terms cancel, giving (for imaginary ϕ, $g_{\phi\bar\phi} = 1$ and $K = 0$)

$$
V = e^{\frac{K}{M_{\rm Pl}^2}}\left[|DW|^2 - \frac{3}{M_{\rm Pl}^2}|W|^2\right]\Big|_{\phi_{\rm can}} = |W'(\sqrt{2}i{\rm Im}\phi)|^2 \equiv \hat{W}'(\phi_{\rm can})^2 \ , \qquad (25.48)
$$

where we have defined the *real function* $\hat{W}$ by

$$
W(\Phi) = \frac{1}{\sqrt{2}}\hat{W}(-\sqrt{2}i\phi). \qquad (25.49)
$$

It follows that we can embed any positive single scalar inflationary potential into the $\mathcal{N} = 1$ supergravity plus chiral superfield scenario. Indeed, if $V \geq 0$, we can take the square root and find $\hat{W}'$, and from it, W. In this way, we can embed new inflation, chaotic inflation (with $V = \lambda\phi^p$), and a lot of other examples.

However, the problem is that the form of K, while looking somewhat simple, is difficult to obtain in string theory, and in this chapter we were not looking at phenomenological supergravity, but rather at the possibility to have a low energy string theory model, via supergravity. From this point of view, the special embedding above is not very good.

Important Concepts to Remember

- Generic Kähler potential and superpotential terms give order one contributions to η.
- The Kähler potential for the volume modulus in KK compactification has the form, for 6 compact dimensions, $K = -3M_{\rm Pl}^2 \ln[\rho + \bar{\rho}]$, which avoids the above constraint.
- Generically, stabilizing the moduli generates quantum corrections that give $\Delta\eta \sim \mathcal{O}(1)$, thus destroying inflation.
- With a vector superfield, we can have D-term inflation, for instance by introducing a FI term to break susy spontaneously.
- One can use field redefinitions and Kähler transformations to restrict the form of possible inflationary models.
- With a canonical Kähler potential $K = \bar{\Phi}\Phi$ and a linear superpotential $W = a\Phi$ we can obtain small ϵ, η, but no end to inflation.
- By adding a subleading term Φ^{n+1}, $n \geq 3$ to W, we get the Izawa-Yanagida model of small field inflation, but here Φ is restricted to remain small.
- We can embed any positive definite potential in supergravity with a chiral superfield using the special embedding with a real function W.

Further reading: See the string inflation reviews [29, 30], and the paper [33].

Exercises

(1) Calculate the inflationary parameters ϵ and η for $\mathcal{N} = 1$ supergravity coupled to a chiral superfield Φ, with Kähler potential

$$K = -3\ln(\Phi + \bar{\Phi}) \tag{25.50}$$

and superpotential

$$W = A\rho. \tag{25.51}$$

(2) Calculate the potential and its equations of motion for the model of $\mathcal{N} = 1$ supergravity coupled to a Super-Maxwell multiplet and 3 chiral superfields $\Phi_\pm$, S, with Fayet-Iliopoulos term $\xi^a V^a$ and superpotential

$$W = S\Phi_+\Phi_- \ , \tag{25.52}$$

and Kähler potential

$$K = |S|^2 + |\Phi_+|^2 + |\Phi_-|^2. \tag{25.53}$$

Here S is Maxwell-neutral and $\Phi_\pm$ have charges ± 1.
(3) Embed chaotic inflation, with $V = \lambda\phi^p$, into $\mathcal{N} = 1$ supergravity coupled to one chiral superfield, via the special embedding.

(4) Consider the $\mathcal{N} = 1$ chiral superfield model with canonical Kähler potential and superpotential

$$K = a\Phi(1 - b\Phi^n). \tag{25.54}$$

Calculate the potential exactly, and then expand it in $\sigma = \mathrm{Re}\ \phi$ to find, in the $n = 3$ and $n = 4$ cases, (25.35) and (25.36), as well as the $n = 1$ and $n = 2$ cases. Repeat the exercise for the Kähler potential $K = -3\ln(\Phi + \bar{\Phi})$.

Chapter 26
Brane (-Antibrane) Inflation

In this chapter we will examine the possibility that we have inflation due to branes, or perhaps a brane-antibrane system. In string theory, we have branes as we saw, in particular Dp-branes. They move in a higher dimensional space (usually 10 dimensions), and then from the point of view of their worldvolume, the positions in the extra dimensions behave as scalars. If moreover we adopt a point of view that our 3+1 dimensional world lives on the worldvolume of these branes (a scenario known as "braneworld"), then the extra dimensions are 3+1 dimensional scalars. It is therefore natural to ask whether one of them, $X(\xi^a)$, can be the inflaton.

26.1 Brane Probe Inflation

If we consider a single brane in flat space, and a single transverse coordinate, the DBI+WZ action takes the form

$$S_{\text{DBI+WZ}} = -T_p \int d^{p+1}\xi \left(\sqrt{1+(\partial_\mu X)^2} - 1\right) , \tag{26.1}$$

where the -1 is the effect of the WZ term in a supersymmetric configuration. As we saw, if we have more than one scalar, the determinant inside the square root can give more terms, but for a single scalar this is all. Expanding at small fields or energies, we find

$$S_{\text{DBI+WZ}} \simeq T_p \int d^{p+1}\xi \left[-\frac{1}{2}(\partial_\mu X)^2 + \frac{1}{8}[(\partial_\mu X)^2]^2 + ...\right] , \tag{26.2}$$

so the action has no potential, but besides the canonical kinetic term, has terms with higher derivatives. This case then doesn't give the standard type of inflation. But a similar case can give inflation from higher derivative kinetic terms, also known as

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_26

k-inflation (after a paper by Armendariz-Picon, Damour and Mukhanov in 1999), as shown in the case of D-branes for instance by Silverstein and Tong in 2003.

However, in a more general background, we can have a potential. In general, the D$p-$brane action is

$$S_p = -T_p \int d^{p+1}\xi e^{-\phi}\sqrt{-\det\left(\partial_a X^\mu \partial_b X^\nu [g_{\mu\nu}(X) + \alpha' B_{\mu\nu}(X)]\right)} + \mu_p \int_{V_{p+1}} C_{(p+1)}. \tag{26.3}$$

For the supersymmetric brane probe, like is the case of the Dp-brane, we have $\mu_p = \pm T_p$, with the sign distinguishing D-branes from anti-D-branes (the sign is anyway relative to a background $C_{(p+1)}$). We need some sort of breaking of supersymmetry in order to have a potential, otherwise the brane probe will not feel a potential. This can be obtained for instance by putting an antibrane in a brane background, or a brane in a background that is not made up only of branes. In those cases, the WZ term will not completely compensate the DBI term, and we will get a potential.

Example

One example is a Dp-brane moving in the background of a doubly-Wick rotated nonextremal Dp-brane (or branes, rather). A nonextremal brane background can be thought of as an extremal brane background (made up of N branes), with added D-brane-anti-D-brane pairs. Indeed, a brane has mass an charge, but in a D-brane-anti-D-brane pair the charges sum up to zero, so we are left with only mass. Then the non-extremal brane background can be thought of as adding mass, i.e., the D-brane-anti-D-brane pairs, to the extremal (D-brane) background.

We make an extremal background with harmonic function $H_p(r)$ nonextremal, by adding a blackening factor (function) $f_p(r)$ as $-dt^2 \to -f_p(r)dt^2$ and $+dr^2 \to +dr^2/f_p(r)$. But moreover, double Wick rotation means that we exchange the roles of t and another worldvolume coordinate x_p, so $-f_p(r)dt^2 + dx_p^2$ becomes $-dt^2 + f_p(r)dx_p^2$.

That means that the solution is

$$\begin{aligned}
ds^2 &= H_p^{-1/2}(r)(-dt^2 + d\vec{x}_{p-1}^2 + f_p(r)dx_p^2) + H_p^{+1/2}(r)\left(\frac{dr^2}{f_p(r)} + r^2 d\Omega_{8-p}^2\right)\\
e^{2\phi} &= g_s^2 H_p^{\frac{3-p}{2}}; \quad B_{\mu\nu} = 0\\
C_{(p+1)} &= \frac{1}{g_s} H_p^{-1} dt \wedge dx^1 \wedge \ldots \wedge dx^p\\
f_p(r) &= 1 - \left(\frac{r_H}{r}\right)^{7-p}\\
H_p(r) &= 1 + \alpha_p \left(\frac{r_p}{r}\right)^{7-p}.
\end{aligned} \tag{26.4}$$

Note that $f_p(r)$ and $H_p(r)$ are harmonic function of the transverse space (r, Ω_{8-p}), as it should be.

We moreover consider the standard static gauge,

$$X^0 = \xi^0, \quad X^i = \xi^i, \quad r = r(\xi^i) = r(X^i). \tag{26.5}$$

Then the induced metric is

$$ds^2_{\text{induced}} = H_p^{-1/2}\left[-(d\xi^0)^2 + \left(\delta_{ij} + \frac{H_p(r)}{f_p(r)}\partial_i r \partial_j r\right) d\xi^i d\xi^j + f_p(r) d\xi_p^2\right]. \tag{26.6}$$

Substituting in the action, we find

$$S = -\frac{T_p}{g_s}\int d^{p+1}\xi H_p^{-1}(r)\left[\sqrt{f_p\left(1 + \frac{H_p(r)}{f_p(r)}\partial_i r \partial_i r\right)} - 1\right]. \tag{26.7}$$

Specialize now to the case $p = 4$, and consider the coordinate $x_p = x_4$ as a KK dimensional reduction coordinate (nothing depends on it), and $\int dx_p = 2\pi R$. Then, by expanding the square root, we obtain the kinetic term

$$-\frac{T_4(2\pi R)}{g_s}\int d^{3+1}\xi \frac{1}{2\sqrt{f_4(r)}}\partial_i r \partial_i r \ , \tag{26.8}$$

and the potential

$$V_4(r) = \frac{T_4 2\pi R}{g_s} H_4^{-1}(r)[\sqrt{f_4(r)} - 1] = +\frac{T_4 2\pi R}{g_s}\frac{1}{1 + \alpha_4\left(\frac{r_4}{r}\right)^3}\left[\sqrt{1 - \frac{r_H^3}{r^3}} - 1\right] < 0. \tag{26.9}$$

Here

$$\begin{aligned} r_4^3 &= \pi g_s N \alpha'^{3/2} \\ \alpha_4 &= \sqrt{1 + \left(\frac{r_H^3}{2r_4^3}\right)^2} - \frac{r_H^3}{2r_4^3}. \end{aligned} \tag{26.10}$$

Note that the potential is negative definite.

We will concentrate in the near extremal case, when $r_H \ll r_4$, so $\alpha_4 \simeq 1$.

It is easy to see that the potential increases monotonically (show that $V_p'(r) > 0$), from $r = r_H$, where

$$V(r_H) = -\frac{T_4 2\pi R}{g_s}\frac{1}{1 + \alpha_4\left(\frac{r_4}{r_H}\right)^3} \simeq -\frac{T_4 2\pi R}{g_s}\left(\frac{r_H}{r_4}\right)^3 \ , \tag{26.11}$$

where in the last equality I used $r = r_H \ll r_4$, and where also

$$V'(r_H) = +\infty \,, \tag{26.12}$$

to $r = \infty$, where $V(r = \infty) = 0$.

For distances $r \gg r_H$, we find

$$V(r) \simeq -\frac{T_4 2\pi R}{2g_s} \left(\frac{r_H}{r}\right)^3 \frac{1 + \frac{1}{4}\left(\frac{r_H}{r}\right)^3}{1 + \alpha_4 \left(\frac{r_4}{r}\right)^3}. \tag{26.13}$$

If moreover $r_4 \gg r \gg r_H$, we find

$$V(r) \simeq -\frac{T_4 2\pi R}{2g_s \alpha_4} \left(\frac{r_H}{r}\right)^3 \left[\frac{r_H^3}{4r^3} - \frac{r^3}{\alpha_4 r_4^3}\right]. \tag{26.14}$$

This potential is approximately constant, and it will give inflation in this region. It is left as an exercise to show that we can satisfy the slow roll conditions.

In conclusion, in this model we have obtained inflation out of the motion of a single D-brane in a nonsupersymmetric background. Other cases of brane motion in backgrounds, without supersymmetry can also result in inflation.

The end of inflation is due to the rapid drop in the potential. As we said in part I, inflation can end either with an oscillation around a minimum, leading to the usual reheating, or with a rapid drop in potential, leading to the nonperturbative preheating. This later case is what happens here, since as we saw $V'(r_H) = +\infty$. As we have described it, it is not obvious, but inflation can end with the inflaton disappearing during this preheating phase. It will become more obvious in the case of *brane-antibrane inflation*.

26.2 Brane-Antibrane Inflation: Flat Space Approximation

Instead of a single brane moving in a background, we can consider a brane and an antibrane attracting each other, and finally annihilating after their collision. In this case, the inflaton will be their *relative* position,

$$\Phi \propto \frac{y}{\alpha'}; \quad y^m = (x_1 - x_2)^m. \tag{26.15}$$

The Dp-brane tension can be expressed as

$$T_p = \alpha' M_s^{p+1} e^{-\phi} \,, \tag{26.16}$$

where M_s is the string mass $1/\sqrt{\alpha'}$. The DBI term in the D-brane action is expanded as

$$S_{\rm DBI} = -\int d^{p+1}x \sqrt{-\gamma}[T_p + \ldots] \,, \tag{26.17}$$

and when considering the brane and antibrane together, the two WZ terms $\pm\mu_p \int C_{(p+1)}$ will cancel out, and the DBI terms will add up to

$$S = -\int d^{p+1}x\sqrt{-\gamma}T_p\left[2 + \frac{1}{4}g_{mn}\gamma^{ab}\partial_a y^m \partial_b y^n + \ldots\right]. \tag{26.18}$$

If the total dimension is written as $D = 4 + n$, the potential energy of the brane-antibrane system per p-dimensional volume is found to be

$$\frac{E}{V_p} = -\beta\left(\frac{e^{2\phi}}{M_s^{2+n}}\right)\frac{T_p^2}{y^{n+1-p}}\ , \tag{26.19}$$

where the calculation (which can be found in the book by Polchiski or in the book by Clifford Johnson on D-branes) gives

$$\beta = -\pi^{-\frac{n+1-p}{2}}\Gamma\left(\frac{n-p+3}{2}\right). \tag{26.20}$$

Define the volume of space in the *extra dimensions along the branes*

$$V_{||} = r_{||}^{p-3}\ , \tag{26.21}$$

and the volume of the coordinates transverse to the branes

$$V_{\perp} = r_{\perp}^{n-p-3}\ , \tag{26.22}$$

we find the Planck scale by identifying the coefficient of the Einstein action in 3+1 dimensions, $M_P^2/2$, with the coefficient of the D-dimensional Einstein action in string theory, $e^{-2\phi}M_s^{2+n}/2$ (so that the Einstein action has the right dimension) after integrating out the extra dimensions, so

$$M_P^2 = e^{-2\phi}M_s^{2+n}V_{\perp}V_{||}. \tag{26.23}$$

In order to trust the calculation, we need to have small α' corrections, which amounts to the volumes $V_{||}$ and $V_{\perp}$ be large in string units, or

$$M_s r_{\perp} \gg 1, \quad M_s r_{||} \gg 1\ , \tag{26.24}$$

which means

$$\frac{1}{r_{\perp}}, \frac{1}{r_{||}} \ll M_s < M_P. \tag{26.25}$$

Given the kinetic action (26.18), the canonically normalized inflaton is

$$\Phi = y\sqrt{\frac{T_p V_{||}}{2}} = y M_s M_P \sqrt{\frac{\alpha' e^{\phi}}{2(M_s r_\perp)^{n+3-p}}}. \tag{26.26}$$

We then find the inflaton potential given by the constant piece in (26.18) and the potential energy (26.19), giving

$$V(y) = A - \frac{B}{y^{n+1-p}}\ , \tag{26.27}$$

where

$$\begin{aligned} A &= 2T_p V_{||} = \frac{2\alpha' e^{\phi}}{(M_s r_\perp)^{n+3-p}} M_s^2 M_P^2 \\ B &= \frac{\beta e^{2\phi}}{M_s^{2+n}} T_p^2 V_{||} = \frac{\alpha^2 \beta e^{\phi} M_P^2}{M_s^{2(n-p+1)} r_\perp^{n-p+3}}. \end{aligned} \tag{26.28}$$

On the other hand, the distance y must be much larger than the string scale, so that we don't have string corrections, but it must also be much smaller than the size of the compact directions $r_\perp$, so that the effect of compactification is small (compactification means that we must impose periodicity on the $r_\perp$ scale, which contradicts the assumptions taken here), thus

$$\sqrt{\alpha'} = M_s \ll y \ll r_\perp. \tag{26.29}$$

For this potential, we can easily find that

$$|\epsilon| \ll |\eta|\ , \tag{26.30}$$

and moreover the expression for η is

$$\eta \simeq -\beta(n-p+2)(n-p+1)\left(\frac{r_\perp}{y}\right)^{n-p+3}. \tag{26.31}$$

Therefore, in order to have small roll inflation, we need $|\eta| \ll 1$, which implies (given that n, p are of order one)

$$y \gg r_\perp. \tag{26.32}$$

But that contradicts the assumption under which we have calculated the potential, so we cannot trust the calculation.

26.3 Brane-Antibrane Inflation in Compact Space

In order to obtain inflation, we must then consider a situation with several branes and one antibrane, or more precisely we consider the correct compactification situation: imposing periodicity in the compact directions means also considering the image branes under the periodicity. The effect is to consider several branes and one antibrane (when the constant piece of the potential is the same, coming from the tension of the probe antibrane) so

$$V(\vec{r}) = A - \sum_i \frac{B}{|\vec{r} - \vec{r}_i|^{n+1-p}}. \tag{26.33}$$

But we find, not surprisingly, that in order to obtain inflation we must be at a special point in the transverse space, namely near an *extremum.*

Square Torus $(\mathbb{R}/\mathbb{Z})^{n+3-p}$

We consider now the simplest possibility, of a square (equal radii, and 90^0 angles) torus in all extra directions. Then by putting the images of a brane in order to take care of the periodicity and obtain the torus from the hyperplane, we find a hyper-cubic lattice, with the hyper-cubic unit cell. The extremum is the center of the hyper-cubic cell. It turns out that at the center, we can put to zero not only the first derivatives of the potential (since it is an extremum, obviously), but also the second derivatives,

$$\begin{aligned}
\left.\frac{\partial V}{\partial r_a}\right|_{\vec{r}=\vec{r}_0} &= (n+1-p)\sum_i \frac{B(r_0 - r_i)_a}{|\vec{r} - \vec{r}_i|^{n+3-p}} = 0 \\
\left.\frac{\partial^2 V}{\partial r_a \partial r_b}\right|_{\vec{r}=\vec{r}_0} &= (n+1-p)\sum_i \frac{B}{|\vec{r}_0 - \vec{r}_i|^{n+3-p}}\left[\delta_{ab} - (n+3-p)\frac{(r_0 - r_i)_a (r_0 - r_i)_b}{|\vec{r}_0 - \vec{r}_i|^2}\right] = 0.
\end{aligned} \tag{26.34}$$

Moreover, just by symmetry, since we are at the center of the cell, we must have all the odd powers of derivatives (giving odd powers of the deviation in the potential) are zero. That means the third derivative terms are also zero, and finally, for $z = r - r_0$,

$$V \simeq A - \frac{1}{4} C z^4 , \tag{26.35}$$

where

$$C = \gamma M_s^{-(2+n)} e^{2\phi} T_p^2 V_{||} r_\perp^{-(n+5-p)}. \tag{26.36}$$

Here γ is a number of order one, and the parametric dependence can be figured out without a precise calculation.

Then one finds (we leave it as an exercise) the slow-roll parameter

$$\eta \simeq -3\gamma \left(\frac{z}{r_\perp}\right)^2. \tag{26.37}$$

Moreover, defining z_* as the z when the scale in the CMBR exits the horizon (and is frozen in), we find

$$z_* = \frac{r_\perp}{\sqrt{2\gamma N_e}} \,, \tag{26.38}$$

where N_e is the corresponding number of e-folds. Then moreover, one gets the relations

$$n_s - 1 = -\frac{3}{N_e}; \quad \frac{dn}{d\ln k} = -\frac{3}{N_e^2}. \tag{26.39}$$

But the inflaton Φ, considered here to be in the regime where $y \gg \sqrt{\alpha'}$, eventually reaches the string scale, where the preheating happens, and then we have to deal with the fact that the interbrane separation (between D-brane and anti-D-brane) gives also rise to a tachyon T in open string theory. This open string tachyon T has the mass squared

$$M_T^2 = \frac{y^2}{(2\pi\alpha')^2} - \frac{2}{\alpha'} \,, \tag{26.40}$$

so it becomes tachyonic when $y \sim \sqrt{\alpha'}$. That signals an instability, which is the fact that when the D-brane and the anti-D-brane collide, they will annihilate (similar to the annihilation of solitons like monopoles, that can be described in terms of annihilation of fundamental particles, with disappearence of charge, but not of mass). In fact, there is a branch of mathematics called K-theory, relevant for these annihilations, which says that one can in fact obtain a brane of lower dimensionality in this annihilation, namely $p \to p - 2$, so that we are consistent with the existence of only odd p in type IIB and even p in type IIA, which is kind of a vortex solution of the worldvolume theory. But the point is that the p-brane charge disappears and mass remains, and can be transferred as energy to various modes.

This generates the reheating (in its preheating form), and it involves many degrees of freedom. In principle, we would need the potential $V(y, T)$, but we know only that at y large T has a large mass, localizing it at $T = 0$ and making it irrelevant, and at small y (in particular $y = 0$), the tachyon dominates instead. But the truncation of the action to the tachyon and specifically the terms with two derivatives, reduced to 4 dimensions, is conjectured to be (supported by many evidences)

$$S_T = -M_s^2 M_P^2 \int d^4x e^{-|T|^2} \left[1 + \kappa_1 \left(\frac{2 + \kappa_2 |T|^2}{M_s^2}\right) |\partial T|^2\right]. \tag{26.41}$$

When canonically normalized, we obtain a Higgs-type potential, and the false vacuum at $T = 0$ is replaced by the true vacuum, of the energy well equal to the tension of the brane (according to a much tested conjecture by Sen). The branes can either result in vacuum, or a D$(p - 2)$-brane, if the tachyon field has a vortex number.

In principle, we can use this action to calculate better (nonperturbatively) the reheating temperature. But a very crude estimate amounts to saying that all the

initial energy of the tachyon will be converted into (reheating) temperature of the modes it gives them to, i.e., that

$$T_{RH}^4 \simeq A \Rightarrow T_{RH} = \left[\frac{2\alpha' e^{\phi}}{(M_s r_{\perp})^{n+3-p}}\right]^{1/4} \sqrt{M_s M_P}. \tag{26.42}$$

We treated here only the example of the antibrane at the center of a square torus with branes, but other examples can be found for antibranes at special points in the transverse space relative to branes. But always, in order to obtain inflation, we must choose a special point, in order to avoid having a quadratic potential (coming from a nonzero second derivative), which would lead to the usual eta problem of string theory. That amounts to a level of fine-tuning of the initial conditions that is hard to explain reasonably, so is is not a very good model.

Important Concepts to Remember

- We get get inflation on a D-brane probe, either of the k-inflation type (only derivatives in the action), or the usual one with a potential.
- An example is a Dp-brane probe in a doubly-Wick rotated background of a nonextremal Dp-brane solution. Inflation ends in preheating, due to a rapid drop in the potential.
- Brane-antibrane inflation in flat space is impossible, meaning that the potential for the brane and antibrane is too steep (cannot be made to inflate).
- In a compact space, considering the infinite number of mirrors for the branes, we obtain a lattice. Starting at a special, extremum, point on the lattice, we can have inflation: the potential is a quartic deviation from a plateau, $V = A - Cz^4/4$.
- Preheating, for the brane-antibrane annihilation, is described by a p-brane action with vortex solutions, describing the K theory process of the annihilation giving a $(p-2)$-brane.

Further reading: See the original paper on brane-antibrane inflation [34], the example of brane inflation in [35].

Exercises

(1) Repeat the calculation of the potential of a Dp-brane probe in a doubly-Wick rotated nonextremal Dp-brane for a Dp'-brane probe, for $p' < p$ and $p' > p$. How does the result change?
(2) Find the conditions for slow roll inflation coming out of the 4-brane in doubly-Wick rotated nonextremal background.
(3) For the case of multi-brane inflation scenario, prove that

$$\eta \simeq -\gamma \left(\frac{z}{r_+}\right)^2, \tag{26.43}$$

and

$$z_* = \frac{r_\perp}{\sqrt{2\gamma N_e}}. \tag{26.44}$$

(4) For the tachyon action in (26.41), find the canonical potential, and write a tachyon vortex ansatz.

Chapter 27
Braneworld Cosmology and the Israel Junction Conditions

In this chapter we will consider the general scenario of a brane moving in an extra dimension, and the generic cosmology it implies. This is a "Brane Universe" scenario, or more properly, *braneworld cosmology*. It is based on the general idea introduced by ADD of large extra dimensions, or by Randall-Sundrum (versions I and II), where the matter (i.e., particle physics) resides on a brane, whereas gravity is the only field that feels the extra dimension. In this scenario, the cosmology is induced by the motion of the brane in the gravitational spacetime with an extra dimension. This is different than the previous incarnation, where the position of a brane acted as a field on the worldvolume (also containing our real 3+1 dimensional world), identified with the inflaton. Here we are mostly interested in "late time" cosmology, in the era that we have radiation and matter.

There are in principle two ways of describing the cosmology:

- from the point of view of the brane: *the brane formalism*, and
- from the point of view of the bulk: *the bulk formalism*.

They are in fact equivalent, so in this chapter I will concentrate on the brane formalism. There is in fact a debate on whether the cosmology described here makes sense, but I will not enter into this discussion, I will only present the tools necessary to understand the models. Although in principle one can consider more dimensions, the analysis becomes very complicated, so we will restrict to a 5 dimensional embedding space. Also, although we are interested in string theory, I should say that this scenario is very general: we only need a brane moving in some higher dimensional space, with a large direction and a negative cosmological constant. However, the details of the construction are not so easy to rigorously obtain in string theory, as is the case for the Randall-Sundrum constructions (I and II).

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_27

27.1 Braneworld Cosmology in the Brane Formalism

We will start with a general 5 dimensional gravity theory with some matter,

$$S = -\frac{1}{2\kappa_{(5)}^2}\int d^5x\sqrt{-G^{(5)}}R^{(5)} + S_{\text{matter}}. \tag{27.1}$$

Here $\kappa_{(5)}^2 = M_{(5)}^{-3}$, where $M_{(5)}$ is the 5-dimensional Planck constant.
We will use a block-diagonal ansatz,

$$ds^2 = G_{AB}dx^A dx^B = g_{\mu\nu}dx^\mu dx^\nu + b^2 dy^2 \ , \tag{27.2}$$

or more generally a *induced metric on the brane*

$$g_{\mu\nu}^{(4)} = G_{\mu\nu}^{(5)} - n_\mu n_\nu \ , \tag{27.3}$$

or more precisely,

$$g_{\mu\nu}^{(4)} = G_{AB} - n_A n_B \ , \tag{27.4}$$

since the components with $A = 5$ vanish. Here n_μ is the normal to the brane.

In the brane formalism adopted here, we will find the Gauss–Codazzi equations and the Israel junction conditions, giving the Einstein's equations on the brane. The same result would be obtained in the bulk formalism, by having a brane moving in a bulk. The latter is usually called "mirage cosmology", which comes from the popular interpretation of "mirage" (which is really an image from something far away) as "something that is not really there, despite appearances". Namely, we have an "illusion of an expanding Universe" generated by the motion of a brane in a geometry. However, both formalisms give the same result.

The principal dynamical tool at our disposal is the 5 dimensional Einstein equations (equations of motion for gravity),

$$R_{AB}^{(5)} - \frac{1}{2}g_{AB}^{(5)}R^{(5)} = \kappa_5^2 T_{AB}. \tag{27.5}$$

Besides it, we need to construct the kinematics of the embedding of the brane. The most important object is the *extrinsic curvature*,

$$K_{\mu\nu} \equiv g_\mu^{(4)\alpha} g_\nu^{(4)\beta}\nabla_\alpha n_b \ , \tag{27.6}$$

where ∇_μ is the 5 dimensional covariant derivative, as opposed to D_μ, which is the 4 dimensional one. We can also define the trace of the extrinsic curvature,

$$K = K^\mu{}_\mu. \tag{27.7}$$

Whereas the intrinsic curvature of an embedded manifold needs no embedding to be defined, the extrinsic one needs it, since it is defined in terms of the normal n_μ and its 5-dimensional covariant derivative. The extrinsic curvature appears also in the Gibbons-Hawking boundary term for the gravitational action, but here we consider a surface inside the manifold.

27.2 Embedding Theory

To describe the embedding of a worldvolume $\{\xi^a\}, a = 0, 1, 2, 3$ into a 5 dimensional spacetime X^A, by $X^\mu = X^\mu(\xi^a)$, $\mu = 0, 1, 2, 3$, we will write down the *Gauss–Codazzi equations* for the embedding.

They are composed of the *Gauss equation*, which in the $5 \to 4$ case is

$$ {}^{(4)}R^\alpha{}_{\beta\gamma\delta} = {}^{(5)}R^\mu{}_{\nu\rho\sigma} g^{(4)\alpha}_\mu g^{(4)\nu}_\beta g^{(4)\rho}_\gamma g^{(4)\sigma}_\delta + K^\alpha{}_\gamma K_{\beta\delta} - K^\alpha{}_\delta K_{\beta\gamma} \,, \tag{27.8} $$

and the *Codazzi equation*, which in the $5 \to 4$ is

$$ D_\nu(K_\mu{}^\nu - K\delta^\nu_\mu) = D_\nu K_\mu{}^\nu - D_\mu K = {}^{(5)}R_{\rho\sigma} n^\sigma g^{(4)\rho}_\mu \,, \tag{27.9} $$

and also the equation

$$ R - K^2 + K_{\mu\nu}K^{\mu\nu} = -2\left(R^{(5)}_{\mu\nu} - \frac{1}{2} G^{(5)}_{\mu\nu} R^{(5)}\right) n^\mu n^\nu \,, \tag{27.10} $$

which however we will not use.

For the metric embedding, we can use (27.3), or rather (27.4). The usual induced metric is defined as

$$ g^{(4)}_{ab} = G^{(5)}_{AB} \partial_a X^A \partial_b X^b \to G^{(5)}_{\mu\nu} \partial_a X^\mu \partial_b X^\nu \,, \tag{27.11} $$

where in the second form we assumed that X^5 is fixed ($\partial_a X^5 = 0$). We can also define

$$ V^A_a = \partial_a X^A \to V^\mu_a = \partial_a X^\mu \,, \tag{27.12} $$

and in the case X^5 fixed we can choose the normal vector to be

$$ n_A = (0, 0, 0, 0, 1) \,, \tag{27.13} $$

such that (valid in general)

$$ G^{(5)}_{AB} n^A n^B = 1 \,, \tag{27.14} $$

and

$$G^{(5)}_{AB} n^A V^B_a = 0. \tag{27.15}$$

Then the four dimensional metric with one up and one down index can be written in alternative forms,

$$g^{(4)\,\alpha}_{\nu} = G^{(5)\mu\alpha} g^{(4)}_{\mu\nu} = \delta^{\alpha}_{\nu} - n^{\alpha} n_{\nu} = V^{\alpha}_{a} V_{a\nu}. \tag{27.16}$$

Substituting this embedding formalism into the 5 dimensional Einstein tensor, we find the relation of the Einstein tensors in 4 and 5 dimensions as

$$\begin{aligned} {}^{(4)}R_{\mu\nu} - \frac{1}{2} g^{(4)}_{\mu\nu}\, {}^{(4)}R = & \left({}^{(5)}R_{\rho\sigma} - \frac{1}{2} g^{(5)}_{\rho\sigma}\, {}^{(5)}R \right) g^{(4)\,\rho}_{\mu} g^{(4)\,\sigma}_{\nu} + {}^{(5)}R_{\rho\sigma} n^{\rho} n^{\sigma} g^{(4)}_{\mu\nu} \\ & + K K_{\mu\nu} - K_{\mu}{}^{\rho} K_{\nu\rho} - \frac{1}{2} g^{(4)}_{\mu\nu} (K^2 - K^{\alpha\beta} K_{\alpha\beta}) - \tilde{E}_{\mu\nu}\,, \end{aligned} \tag{27.17}$$

where

$$\tilde{E}_{\mu\nu} = {}^{(5)}R^{\alpha}{}_{\beta\rho\sigma} n_{\alpha} n^{\rho} g^{(4)\,\beta}_{\mu} g^{(4)\,\sigma}_{\nu}. \tag{27.18}$$

This is left as an exercise to prove.

Next we decompose the 5 dimensional Riemann tensor into a Ricci part (tensor and scalar) and Weyl tensor. The Weyl tensor is an object that is invariant under conformal rescalings (rescalings of the metric by a conformal factor). That means that in particular, it is zero in the case of conformally flat metrics, like the AdS case. This will be useful, since as we will see, the bulk will in fact be of the AdS type, so only fluctuations away from it will have nonzero Weyl tensor.

The definition of the Weyl tensor in a general dimension D is given by

$$\begin{aligned} {}^{(D)}C_{\alpha\beta\gamma\delta} = {}^{(D)}R_{\alpha\beta\gamma\delta} - \frac{2}{D-2} \left(g_{\alpha[\gamma}\, {}^{(D)}R_{\delta]\beta} - g_{\beta[\gamma}\, {}^{(D)}R_{\delta]\alpha} \right) + \frac{2}{(D-1)(D-2)}\, {}^{(D)}R\, g_{\alpha[\gamma} g_{\delta]\beta}. \end{aligned} \tag{27.19}$$

Inverting this relation in 5 dimensions, we write

$$ {}^{(5)}R_{\alpha\beta\gamma\delta} = {}^{(5)}C_{\alpha\beta\gamma\delta} + \frac{2}{3} \left(G_{\alpha[\gamma}\, {}^{(5)}R_{\delta]\beta} - G_{\beta[\gamma}\, {}^{(5)}R_{\delta]\alpha} \right) + \frac{1}{6}\, {}^{(5)}R\, G_{\alpha[\gamma} G_{\delta]\beta}. \tag{27.20}$$

Substituting this decomposition into the above relation for the 4 dimensional Einstein tensor in terms of the 5 dimensional Einstein tensor, and using the 5 dimensional Einstein's equations (27.5), we find

$$\begin{aligned} {}^{(4)}R_{\mu\nu} - \frac{1}{2} g^{(4)}_{\mu\nu}\, {}^{(4)}R = & \frac{2\kappa^2_{(5)}}{3} \left[T_{\rho\sigma} g^{(4)\,\rho}_{\mu} g^{(4)\,\sigma}_{\nu} + \left(T_{\rho\sigma} n^{\rho} n^{\sigma} - \frac{1}{4} T^{\rho}{}_{\rho} \right) g^{(4)}_{\mu\nu} \right] \\ & + K K_{\mu\nu} - K_{\mu}{}^{\sigma} K_{\nu\sigma} - \frac{1}{2} g^{(4)}_{\mu\nu} (K^2 - K^{\alpha\beta} K_{\alpha\beta}) - E_{\mu\nu}\,, \end{aligned} \tag{27.21}$$

where now $E_{\mu\nu}$ is a traceless object depending only on the Weyl tensor, so vanishing in pure AdS space

$$E_{\mu\nu} \equiv {}^{(5)}C^{\alpha}{}_{\beta\rho\sigma} n_\alpha n^\rho g^{(4)\beta}_{\mu} g^{(4)\sigma}_{\nu}; \quad E^{\mu}{}_{\mu} = 0. \tag{27.22}$$

From the Codazzi equation (27.9) and the 5 dimensional Einstein's equations (27.5), we find

$$D_\nu K_\mu{}^\nu - D_\mu K = \kappa^2_{(5)} T_{\rho\sigma} n^\sigma g^{(4)\rho}_{\mu}. \tag{27.23}$$

27.3 Braneworld and Cosmology

In order to describe braneworld cosmology, we will assume that the energy-momentum tensor has a component on the brane, and a bulk component that is simply a (negative) bulk cosmological constant, namely

$$T_{\mu\nu} = -\Lambda g_{\mu\nu} + S_{\mu\nu} \frac{\delta(y)}{b}. \tag{27.24}$$

Moreover, we will assume that the brane energy-momentum tensor has also a brane cosmological constant component λ, and a matter component $\tau_{\mu\nu}$, so

$$S_{\mu\nu} \equiv -\lambda g^{(4)}_{\mu\nu} + \tau_{\mu\nu}. \tag{27.25}$$

It is important that we have $\lambda \neq 0$. In fact, in the first paper on the subject, λ was not considered, and then one obtained the non-traditional Friedman equation

$$H^2 = \left(\frac{\rho_t}{6M^3_{(5)}} \right)^2 + \frac{\Lambda_{\text{bulk}}}{6M^3_{(5)}}\,, \tag{27.26}$$

which does not reduce to the usual cosmology, since there we have $H^2 \propto \rho$ instead. Here ρ_t is the total energy density. It was quickly realized however what the solution is. Consider that the energy density has a cosmological constant component that cancels the bulk one, i.e., that

$$\rho_t = \lambda + \rho\,, \tag{27.27}$$

and that

$$\Lambda_{\text{bulk}} = -\frac{\lambda^2}{6M^3_{(5)}}. \tag{27.28}$$

Then, one gets the Friedmann equation

$$H^2 = \pm \frac{\rho_\pm}{3M^2_{\mathrm{Pl},(4)}} \left(1 \pm \rho_\pm \frac{M^2_{\mathrm{Pl}(4)}}{16M^6_{(5)}}\right) , \tag{27.29}$$

which is just the usual one, $H^2 \propto \rho$, just with a ρ^2 correction term.

Note that this is all in the case that the matter component of the brane energy-momentum tensor is the usual diagonal form,

$$\tau_{\mu\nu} = \mathrm{diag}(-\rho, P, P, P) \Rightarrow T_{\mu\nu,\mathrm{brane}} = \frac{\delta(y)}{b}\mathrm{diag}(-\rho, P, P, P). \tag{27.30}$$

For a *static* fifth dimension, the radius of the extra dimenion, b, is constant, $\dot{b} = 0$, so we can choose $b = 1$. Note that b, appearing as $G_{55} = b^2$, is the radius of the extra dimension, or in terms of the 4 dimensional field theory, a *radion*. This condition of $\dot{b} = 0$ means that the *radion is stabilized*. This is in fact needed, since otherwise we would observe a fifth scalar force; this is the same argument we used to say that all the moduli of the string compactification must be stabilized.

But for the radion of a fifth dimension, there is a very simple mechanism for stabilization that was found by Goldberger and Wise. One adds a scalar field Φ, with bulk action

$$S_{\mathrm{bulk}} = \int d^5x\sqrt{-G}\left(-\frac{1}{2}G^{AB}\partial_A\Phi\partial_B\Phi - \frac{m^2}{2}\Phi^2\right) , \tag{27.31}$$

as well as a potential term *on the brane*, of the usual Higgs (Mexican hat) type,

$$S_{\mathrm{brane}} = -\int d^4x\sqrt{-g^{(4)}}\lambda(\Phi^2 - v^2)^2. \tag{27.32}$$

This is in fact enough to stabilize the radion (the radius of the fifth -extra-dimension), so $\dot{b} = 0$.

27.4 Israel Junction Conditions

The Israel junction conditions are the equivalent of the Einstein's equations *on the brane*. In fact, even though strictly speaking they are postulated, they can be argued for ("derived") by considering the (second order) Einstein's equations with a delta function in the source (energy-momentum tensor), and integrating them in an infinitesimal region around the brane. We can use then Codazzi's equation and the Einstein equations to obtain the Israel conditions. This is a familiar trick: When having some second order differential equation with a delta function term (for instance, for the Schrödinger equation in one dimension, with a delta function potential), we integrate

the equations, and obtain a relation for the "jump" in the first derivative, with respect to the coefficient of the delta function term.

Using the notation

$$[f(x)] \equiv \lim_{\epsilon\to 0}[f(x+\epsilon) - f(x-\epsilon)] \equiv \lim_{x\to 0+} f(x) - \lim_{x\to 0-} f(x) \equiv f^+ - f^- \,, \tag{27.33}$$

the Israel junction conditions are

$$[g^{(4)}_{\mu\nu}] = 0 \,, \tag{27.34}$$

i.e., the 4 dimensional metric is continuous across the surface, and

$$[K_\nu{}^\mu - K\delta^\mu_\nu] = -\kappa^2_{(5)} S^\mu_\nu . \tag{27.35}$$

The latter can be rewritten as (taking the trace and replacing)

$$[K_{\mu\nu}] = -\kappa^2_{(5)}\left(S_{\mu\nu} - \frac{1}{3} g^{(4)}_{\mu\nu} S\right) \tag{27.36}$$

Moreover, we can make a standard assumption in braneworld physics, used to simplify the models, namely the assumption of $\mathbb{Z}_2$ symmetry of the physics with respect to the brane, which gives

$$K^+_{\mu\nu} = -K^-_{\mu\nu} = -\frac{\kappa^2_{(5)}}{2}\left(S_{\mu\nu} - \frac{1}{3} g^{(4)}_{\mu\nu} S\right). \tag{27.37}$$

Then, substituting (27.37) into (27.21), we obtain the gravitational (Einstein's) equations on the 3-brane, in the form

$${}^{(4)}R_{\mu\nu} - \frac{1}{2} g^{(4)}_{\mu\nu}{}^{(4)}R = -\Lambda_4 g^{(4)}_{\mu\nu} + 8\pi G_N \tau_{\mu\nu} + \kappa^2_{(5)}\pi_{\mu\nu} - E_{\mu\nu} \,, \tag{27.38}$$

where

$$\Lambda_4 = \frac{\kappa^2_{(5)}}{2}\left(\Lambda + \frac{\kappa^2_{(5)}\lambda^2}{6}\right) \tag{27.39}$$

is the residual (4 dimensional) cosmological constant, if the bulk Λ and brane λ don't cancel each other. From the coefficient of the energy-momentum tensor $\tau_{\mu\nu}$ we find the Newton's constant as

$$G_N = \frac{\kappa^4_{(5)}\lambda}{48\pi} \,, \tag{27.40}$$

and there is an extra term, quadratic in the energy-momentum tensor (giving the ρ^2 term advertised before),

$$\pi_{\mu\nu} = -\frac{1}{4}\tau_{\mu\alpha}\tau_{\nu}{}^{\alpha} + \frac{\tau}{12}\tau_{\mu\nu} + \frac{1}{8}g^{(4)}_{\mu\nu}\tau_{\alpha\beta}\tau^{\alpha\beta} - \frac{1}{24}g^{(4)}_{\mu\nu}\tau^2. \tag{27.41}$$

The energy-momentum tensor $\tau_{\mu\nu}$ satisfies the usual conservation equation. Indeed, from (27.23) and (27.37) we find that

$$D_\nu K_\mu{}^\nu - D_\mu K \propto D_\nu \tau_\mu{}^\nu = 0. \tag{27.42}$$

Here $E_{\mu\nu}$ is the same as before, but cannot be freely specified. Indeed, from the 4 dimensional Bianchi identity

$$D^\mu \left({}^{(4)}R_{\mu\nu} - \frac{1}{2}g^{(4)}_{\mu\nu}{}^{(4)}R \right) = 0 \,, \tag{27.43}$$

substituted in the equations of motion (27.38) and using the conservation equation for $\tau_{\mu\nu}$, we find

$$\begin{aligned} D^\mu E_{\mu\nu} &= K^{\alpha\beta}(D_\nu K_{\alpha\beta} - D_\beta K_{\nu\alpha}) \\ &= \frac{\kappa^2_{(5)}}{4}\left[\tau^{\alpha\beta}(D_\nu\tau_{\alpha\beta} - D_\beta\tau_{\nu\alpha} + \frac{1}{3}(\tau_{\mu\nu} - g^{(4)}_{\mu\nu}\tau)D^\mu\tau\right]. \end{aligned} \tag{27.44}$$

27.5 FLRW Cosmology

As advertised, within the context of FLRW cosmology, with 4 dimensional metric

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu = -dt^2 + a^2(t)g^{(3)}_{ij}dx^i dx^j \,, \tag{27.45}$$

and for the usual diagonal energy-momentum tensor

$$\tau^\mu{}_\nu = \mathrm{diag}(-\rho(t), P(t), P(t), P(t)) \,, \tag{27.46}$$

we obtain the modified Friedmann equations

$$\begin{aligned} H^2 + \frac{\kappa}{a^2} &= \frac{\Lambda_4}{3} + \frac{8\pi G_N}{3}\rho\left(1 + \frac{\rho}{2\lambda}\right) + \frac{\mu}{a^4} \\ \dot{H} - \frac{\kappa}{a^2} &= -4\pi G_N(\rho + P)\left(1 + \frac{\rho}{\lambda}\right) - \frac{2\mu}{a^4}. \end{aligned} \tag{27.47}$$

Here $\kappa = 0, \pm 1$ characterizes as usual the curvature (flat, closed, open) of the 3 dimensional Universe, but we notice the appearance of a "dark radiation" term μ/a^4. The $1/a^4$ behaviour is one of radiation, but the constant μ comes as an integration constant, when integrating the constraint (27.44), a fact that is left as an exercise to prove.

As we mentioned, the bulk formalism gives the same result, even though the methods seem different. There, we consider a moving brane in a static bulk ("mirage cosmology"), and we solve for the bulk geometry (solve the 5 dimensional Einstein equations). After that, we solve for the brane motion, with energy-momentum tensor on the brane, in this fixed geometry and impose the Israel junction conditions. Together with conditions from the bulk this will again lead to the same modified Friedmann equations.

Important Concepts to Remember

- Braneworld cosmology appears as the motion of a brane in a spacetime, and is described either in the brane formalism or the bulk formalism. It primarily describes "late time" cosmology, close to the RD and MD regimes.
- In the bulk formalism, it is called "mirage cosmology"; we use the brane formalism.
- The embedding of a brane in a bulk is defined by the Gauss–Codazzi equations (the Gauss equation and the Codazzi equation), depending on the Riemann curvature tensor and the extrinsic curvaturre.
- For braneworld cosmology, we must assume a negative cosmological constant in the bulk Λ and on the brane λ (as a part of the brane energy-momentum tensor $\tau_{\mu\nu}$), related by $\Lambda \propto -\lambda^2$, like in the RS model.
- Then, instead of $H^2 \propto \rho^2 + \Lambda$, we get a linear term, $H^2 \propto \rho(1 + \rho)$, with a small correction to the usual cosmology.
- A static fifth dimension can be achieved by stabilizing the radion via the Goldberger-Wise mechanism, with a massive bulk scalar field with a Higgs potential on the brane.
- The Israel junction conditions are equivalent to the Einstein's equations on the brane, and related the jump in bulk quantities with at most one derivative with the delta function source on the brane, via integration of the (second order) Einstein equations.
- The Israel junction conditions are $[g^{(4)}_{\mu\nu}] = 0$ and $[K_\nu{}^\mu - K\delta^\mu_\nu] = -\kappa_5^2 S_\nu{}^\mu$, with $S_\nu{}^\mu$ the energy-momentum tensor on the brane, including its cosmological constant λ.
- One obtains a corrected FLRW cosmology, with dark radiation appearing as an integration constant.

Further reading: The original braneworld cosmology, with $H^2 \propto \rho^2$, appeared in [36]. Almost immediately, the correction introduced by the compensating brane cosmological constant, obtaining $H^2 \propto \rho + \rho^2$, was found in [37, 38]. The Goldberger-Wise radion stabilization mechanism was presented in [39]. Taking this into account, the cosmology with $\dot{b} = 0$ was considered in [40] and especially [41], which is now the standard treatment, which I mostly followed. Reviews including braneworld cosmology are [42], which is actually an early review on string cosmology, and mentions braneworld cosmology in this context, and the general modified gravity and cosmology review [43], which contains a comparison of the brane and bulk formalisms, and cosmological perturbations in them. The ADD scenario was defined in [44], extended (via strings) in [45], and the two Randall-Sundrum scenarios were defined in [46,

47]. The validity of the Israel junction conditions, as coming from the Einstein's equations, was studied in [48].

Exercises

(1) Show that the Gauss–Codazzi equations imply

$$
\begin{aligned}
{}^{(4)}R_{\mu\nu} - \frac{1}{2} g^{(4)(4)}_{\mu\nu} R = & \left({}^{(5)}R_{\rho\sigma} - \frac{1}{2} g^{(5)(5)}_{\rho\sigma} R \right) g^{(4)\rho}_{\mu} g^{(4)\sigma}_{\nu} + {}^{(5)}R_{\rho\sigma} n^{\rho} n^{\sigma} g^{(4)}_{\mu\nu} \\
& + K K_{\mu\nu} - K_{\mu}{}^{\rho} K_{\nu\rho} - \frac{1}{2} g^{(4)}_{\mu\nu} (K^2 - K^{\alpha\beta} K_{\alpha\beta}) - \tilde{E}_{\mu\nu} \, ,
\end{aligned}
\tag{27.48}
$$

where

$$
\tilde{E}_{\mu\nu} = {}^{(5)}R^{\alpha}{}_{\beta\rho\sigma} n_{\alpha} n^{\rho} g^{(4)\beta}_{\mu} g^{(4)\sigma}_{\nu} . \tag{27.49}
$$

(2) Derive (argue for) the Israel junction conditions from integrating the Einstein's equations around the brane.
(3) Find the modified Friedman equations with μ/a^4 integration constant term from the integration of the equation for $D^{\mu} E_{\mu\nu}$.
(4) Find the solution for $a(t)$ solving the modified Friedmann equations (27.47) for a flat Universe ($\kappa = 0$), given a scale factor a_0 at time t_0, with density ρ_0 and pressure P_0.

Chapter 28
The KKLT Scenario for de Sitter Backgrounds in String Theory

In this chapter we will learn how to obtain de Sitter backgrounds in string theory. This is a difficult problem. In particular, there was a "no-go theorem" about the non-existence of de Sitter backgrounds in the supergravity approximation of string theory, by Juan Maldacena and Carlos Núñez. Kachru, Kallosh, Linde and Trivedi (KKLT) have shown for the first time that it is in fact possible, with the introduction of some nonperturbative string information.

However, it is worth noting that recently there was a lot of controversy on whether this is truly possible, in particular articles by I. Bena and collaborators claming the approximations used in the calculations are not valid, and there is no true de Sitter vacuum and, even more recently, the "swampland conjecture" of Obied, Ooguri, Spodyneiko and Vafa, claiming that there is a string theory version of the no-go theorem that precludes de Sitter vacua, and we can have only a certain type of metastable situations, leading to quintessence models.

28.1 Supersymmetric Vacua Are AdS

Let us prove that in supergravity, the constraint of having supersymmetry for the vacuum implies it is of AdS type ($\Lambda < 0$), not dS type.

The scalar potential coming out of supergravity coupled to an $\mathcal{N} = 1$ chiral superfield and an $\mathcal{N} = 1$ vector superfield is

$$V = e^{K/M_{\rm Pl}^2}\left(g^{\alpha\bar{\beta}}D_\alpha W\overline{(D_\beta W)} - \frac{3}{M_{\rm Pl}^2}W\right) + \frac{1}{2}(\mathrm{Re}\, f)^{-1\ AB}D_A D_B\ , \tag{28.1}$$

where on the first line we have the F-term from the chiral superfield, and on the second we have the D-term from the vector superfield, with f the gauge kinetic function.

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_28

The condition for a supersymmetric vacuum is, as we said, $Q|0\rangle = 0$, which in terms of fields means that the variation of the fields in the background must be zero, $\delta_{\rm SUSY}$ fields$=0$. But the variation of a boson is a fermion, which has zero VEV due to Lorentz invariance, so we only need to impose the vanishing of the variation of the fermions (which is a bosonic field). The variation of the spin 1/2 fields is as follows. The variation of the fermion in the chiral multiplet is

$$\delta_{\rm SUSY}\chi^{\alpha} = -\frac{1}{\sqrt{2}} e^{\frac{K}{2M_{\rm Pl}^2}} g^{\alpha\bar{\beta}} \overline{D_{\beta} W}. \tag{28.2}$$

On the other hand, the variation of the gaugino (the fermion in the vector multiplet) is

$$\delta_{\rm SUSY} P_L \lambda^A = \frac{i}{2} (\mathrm{Re}\, f)^{-1AB} D_B \ , \tag{28.3}$$

where A, B is an index in the adjoint of the gauge group. We therefore see that the conditions for supersymmetry are

$$D_{\alpha} W = 0; \quad D_A = 0. \tag{28.4}$$

The condition on the chiral superfield in rigid susy (when $1/M_{\rm Pl}^2 \to 0$) is indeed $\partial_{\alpha} W = 0$, and the condition for the vector is $D_A = 0$.

But what about the gravitino? The variation of the gravitino is

$$\delta_{\rm SUSY} P_L \psi_{\mu} = \frac{1}{2M_{\rm Pl}^2} e^{\frac{K}{2M_{\rm Pl}^2}} W \gamma_{\mu} \ , \tag{28.5}$$

and we note that it is not strictly speaking linear in the fields: it has a constant piece. In fact, when putting the fields in their vacuum, the right hand side will become a constant (W is constant), which signifies the presence of the AdS space, as we will see. In the vacuum, we obtain

$$V = -3 \frac{|W|^2}{M_{\rm Pl}^2} e^{\frac{K}{M_{\rm Pl}^2}} < 0. \tag{28.6}$$

So indeed we have an AdS background, associated with a constant W.

In order to obtain a de Sitter background, we therefore need to break supersymmetry. But the breaking must be done in a *controlled* way, so that we can calculate. We will be doing it by adding antibranes in a background generated by branes (with charges opposite to the ones of the background).

28.2 General Set-up for KKLT

The general set-up that we will use for the KKLT scenario is of type IIB string theory flux compactification, with some corrections.

More precisely, we will treat type IIB string theory compactified on a CY_3 space, *with varying* $\tau = a + ie^{-\phi}$. Note that for this case there is a fancy name: it is called "F-theory on CY_4". Unlike M-theory, which is string theory at strong coupling, but is genuinely 11 dimensional: there is dynamics in the 11th dimension, as well as Lorentz invariance, in the F-theory case, this is just a mathematical trick. We write the theory as a 12 dimensional theory with signature $(10, 2)$, but we have no dynamics of the extra dimensions (fields don't depend on the 2 extra coordinates), nor Lorentz invariance. Then τ acts as the modulus of a torus (if τ is constant) on which we compactify, or (if τ varies) as the torus fiber of a fibration of a CY_4 space with a CY_3 base. We also have a G-flux, with

$$G_3 = F_3 - \tau H_3 \tag{28.7}$$

being nonzero on the compact space. As we said, we have the GVW superpotential

$$W = \int_M G \wedge \Omega. \tag{28.8}$$

The volume modulus on the CY_3 is

$$\rho = \frac{b}{\sqrt{2}} + ie^{4u}\ , \tag{28.9}$$

where b comes from a B-field, and the metric on the compact space is $g_{mn} \propto e^{2u}$.

The Kähler potential for the moduli ρ, τ, Ω, where Ω is the (unique) holomorphic 3-form on the CY_3 is

$$\frac{K}{M_{\rm Pl}^2} = -3\ln[-i(\rho - \bar{\rho})] - \ln[-i(\tau - \bar{\tau})] - \ln\left[-i\int_M \Omega \wedge \bar{\Omega}\right]. \tag{28.10}$$

If W is independent of ρ (like in the GVW superpotential above), then

$$\begin{aligned} D_\rho W &= -\frac{3M_{\rm Pl}^2}{(\rho - \bar{\rho})} W \\ g_{\rho\bar{\rho}} &= -\frac{3M_{\rm Pl}^2}{(\rho - \bar{\rho})^2} \\ g^{\rho\bar{\rho}} D_\rho W \overline{D_\rho W} &= \frac{3}{M_{\rm Pl}^2}|W|^2. \end{aligned} \tag{28.11}$$

That means that in the potential, the constant piece $|W|^2$ cancels against the ρ contribution, so

$$V = e^{\frac{K}{M_{\rm Pl}^2}} \left[\sum_{a,b} g^{a\bar{b}} D_a W \overline{D_b W} - \frac{3}{M_{\rm Pl}^2} |W|^2 \right]$$
$$= e^{\frac{K}{M_{\rm Pl}^2}} \left[\sum_{i,j} g^{i\bar{j}} D_i W \overline{D_j W} \right] \geq 0 \,, \tag{28.12}$$

where a, b are all the moduli, and i, j are a, b less ρ. Note that now the potential is positive definite! So it admits a zero energy supersymmetric vacuum (Minkowski), though of course we still cannot have a de Sitter vacuum.

But in reality, while there are no *perturbative* ρ contributions to the superpotential (since the volume is perturbatively a modulus), there are *nonperturbative* contributions to it, coming from two sources:

(1) Euclidean D3-brane "instantons". These are extensions of the usual field theoretic instantons, which are Euclidean solutions localized both in space and time in 3+1 dimensions. Now we consider Euclidean signature D3-branes wrapping 4-cycles in the geometry, i.e., all of their coordinates (even the Euclidean "worldvolume time") compact, leaving an instanton (localized in space and time) in our 3+1 dimensions.

Then, as we saw, $g_{mn} \propto e^{2u}$, where the volume of the 4-cycle is $\sqrt{g} \propto e^{4u} = \text{Im}\,\rho$. Then the contribution $e^{iS} \sim e^{iVol}$ (the action for the wrapped D3-brane is proportional to the volume of the 4-cycle) becomes a contribution to the superpotential,

$$W_{\rm inst.} = T(z_i) e^{2\pi i \rho} \,, \tag{28.13}$$

where z_i are the complex structure moduli.

(2) Gluino condensation in some $\mathcal{N} = 1$ supersymmetric $SU(N_c)$ gauge theory on D7-branes wrapping 4-cycles.

If we have D7-branes, wrapped on a 4-cycle on the internal manifold, they will still be extended in the 3+1 noncompact directions, and have a $SU(N_c)$ gauge theory on them. But in a supersymmetric gauge theory we can have spontaneous supersymmetry breaking, appearing through the condensation of the gluinos λ, namely a nonzero VEV for a gluino bilinear $\langle \bar{\lambda}\lambda \rangle$. The superpotential responsible for this is $W = \Lambda_{N_c}^3$, where Λ_{N_c} is the dynamically generated scale, of the type $e^{\frac{iS_i}{N_c}}$. Here the S_i is the instanton-type action, which, since $g_{YM}^2 = 4\pi g_s$ on the D3-brane (relation between the open and closed string couplings) and $g_{YM}^2 = \frac{4\pi g_s}{R^4}$ on D7-brane, gives

$$\frac{8\pi^2}{g_{YM}^2} = 2\pi \frac{R^4}{g_s} = 2\pi \text{Im}\rho. \tag{28.14}$$

Here we have used the exact relation $\text{Im}\ \rho = e^{4u} = R^4/g_s$ (since e^{2u} plays the role of R^2, as common factor of the compact metric). Finally, we have

$$W_{\text{gluino}} = Ae^{\frac{2\pi i\rho}{N_c}}. \tag{28.15}$$

Together, the two nonperturbative sources above are generically written as

$$W = Ae^{ia\rho}. \tag{28.16}$$

On the other hand, the perturbative superpotential was independent of ρ, so we will describe it as the constant W_0. In total, we have the superpotential

$$W = W_0 + Ae^{ia\rho}. \tag{28.17}$$

28.3 Moduli Stabilization in KKLT

In general, flux compactifications have the property that the string coupling g_s, and by extension τ, as well as the complex structure (shape) moduli are stabilized (fixed) at a scale $m \sim \frac{\alpha'}{R^3}$ (it is proportional to α', and by dimensional analysis, must also be proportional to $1/R^3$).

To understand this, we take the example of a flux of the usual electromagnetic field $F_{\mu\nu}$ on S^2, which is quantized,

$$\int_{S^2} F_{\mu\nu} dx^\mu \wedge dx^\nu = N. \tag{28.18}$$

The reason for this is the *Dirac quantization condition* in 4 dimensions. The electric charge e and magnetic charge g can be defined as

$$\int_{S^2} *F = e; \quad \int_{S^2} F = g. \tag{28.19}$$

Then the Dirac quantization condition, obtained by considering the quantum mechanics invariance under gauge transformations (we will not explain this further here) is

$$\frac{1}{2\pi} eg \in \mathbb{Z}. \tag{28.20}$$

The result of the quantization condition is that, if *a single magnetic monopole exists in the Universe*, then electric charge is quantized. Since we know experimentally that electric charge seems quantized, it seems very likely that there is a magnetic monopole somewhere.

Reversely, the existence of a single electric charge (which we know to be true) implies that $\int_{S^2} F$ is quantized, as we said. Then, it follows that the solution for the electromagnetic field is

$$F_{\mu\nu} = \frac{N}{4\pi R^2} \epsilon_{\mu\nu}. \tag{28.21}$$

But in turn that means that, if we now *deform* the 2-sphere, it will cost energy (since it will deform the relation above), which means that the complex structure (shape) moduli will be stabilized.

The quantization condition was generalized to p-forms by Nepomechie and Teitelboim. Defining

$$\mu_p = \int *F_{p+2} \ , \tag{28.22}$$

in 10 dimensions the generalized quantization condition is

$$\frac{1}{2\pi}\mu_p\mu_{6-p} \in \mathbb{Z}. \tag{28.23}$$

That means that all the moduli are stabilized, except ρ (in fact, also the Kähler moduli different than ρ are stabilized). We will not describe the details here. We can therefore ignore all other moduli, and concentrate on K and W for ρ. The Kähler potential is, as we saw

$$K = -3\ln[-i(\rho-\bar{\rho})] \ , \tag{28.24}$$

and the superpotential is

$$W = W_0 + Ae^{ia\rho}. \tag{28.25}$$

Here $a \in \mathbb{R}_+$, and W_0, $A \in \mathbb{R}$, and we can also take $W_0 < 0$. We will also consider the case when there is no nonzero axion, so $\rho = i\sigma$.

In a supersymmetric vacuum, as we saw, we need to put $D_\rho W = 0$. But now

$$D_\rho W = iAae^{ia\rho} - \frac{3}{\rho-\bar{\rho}}(W_0 + Ae^{ia\rho}) \ , \tag{28.26}$$

and on $\rho = i\sigma$, we obtain

$$D_\rho W|_{\rho=i\sigma} = i\left[Aae^{-a\sigma} + \frac{3}{2\sigma}(W_0 + Ae^{-a\sigma})\right]. \tag{28.27}$$

Moreover, on $\rho = i\sigma$, we also have

$$\begin{aligned} K|_{\rho=i\sigma} &= -2\ln 2\sigma \\ W|_{\rho=i\sigma} &= W_0 + Ae^{-a\sigma} \\ g_{\rho\bar{\rho}}|_{\rho=i\sigma} &= -\left.\frac{3}{(\rho-\bar{\rho})^2}\right|_{\rho=i\sigma} = \frac{3}{4\sigma^2}. \end{aligned} \tag{28.28}$$

Then the supersymmetric vacuum condition $D_\rho W|_{\rho=i\sigma_0}$ becomes

$$W_0 = -Ae^{-a\sigma_0}a\left(1 + \frac{2a\sigma_0}{3}\right). \tag{28.29}$$

The potential at the minimum is then

$$V(\rho = i\sigma_0) = e^{K/M_{\rm Pl}^2}\left(|D_\rho W|^2 - \frac{3}{M_{\rm Pl}^2}|W|^2\right)\Big|_{\rho=i\sigma_0} = -\frac{3}{M_{\rm Pl}^2}|W|^2 e^{K/M_{\rm Pl}^2}|_{\rho=i\sigma_0}$$
$$= -\frac{A^2a^2}{6\sigma_0}e^{-2a\sigma_0} < 0\,, \tag{28.30}$$

where in the last equality we have used the condition at the minimum (28.29). We see that again, the minimum is of AdS type.

In order to trust the calculation above, we must have small worldsheet string α' corrections, which means that the compactification volume must be large in α' units, or $\sigma \gg 1$. Moreover, we have used a single exponential $Ae^{ia\rho}$, but at least in the instanton calculation, there could be an infinite series of instantons of the type $A_n e^{ian\rho}$. In order for the leading instanton to be a good approximation, we need $a\sigma > 1$. Together, these two conditions are satisfied by

$$a < 1, \quad \sigma \gg 1, \quad a\sigma_0 > 1. \tag{28.31}$$

Moreover, we can then *choose* $W_0 \ll 1$, since this is now consistent with (28.29), independently of A. We will see that we need this condition for the de Sitter vacua.

28.4 Compactification

We now give a few details on the compactification. A bit more will be given in the next chapter. The background metric (in Einstein frame) will be of the type

$$ds^2_{10,E} = e^{2A(y)}\eta_{\mu\nu}dx^\mu dx^\nu + e^{-2A(y)}\tilde{g}_{mn}(y)dy^m dy^n\,, \tag{28.32}$$

and it will be a solution of the type IIB equations of motion known as "Klebanov–Strassler" (KS), after its discoverers.

The solution has a radial coordinate $y = \sqrt{\sum_n (y^n)^2}$ and as a function of it, the volume of the extra coordinates changes from a minimum value to $+\infty$, forming what is known as a (semi-infinite) "cigar" geometry. But instead of this infinite (decompactified) geometry, we cut the KS geometry at some $y_{\rm max}$, and glue onto it a half of a CY_3 geometry, in order to make it compact. This can be done smoothly, so that we still have a solution of the Einstein's equations. Moreover, then we add fluxes in order to stabilize the moduli.

In this construction then, there is a minimum value for the volume, i.e., a minimum value for $A(y)$, giving

$$a_{\rm min} = e^{A_{\rm min}} \sim e^{-\frac{2\pi K}{3g_s M}} \ll 1. \tag{28.33}$$

Here M, K are integers defining the fluxes through the geometry, and we can choose them such as to have $a_{\min} \ll 1$.

28.5 de Sitter Vacua

We finally come to the construction of de Sitter vacua. Since the background has D-brane charges, a way to break supersymmetry is to introduce anti-D-branes. Specifically, we will introduce an anti-D3-brane in the geometry, at position y_0. Then, if

$$a_0 \equiv e^{A(y_0)} \,, \tag{28.34}$$

the anti-D3-brane adds the extra energy density to the system of

$$\delta V = \frac{2a_0^4 T_3}{g_s^4} \frac{1}{(\mathrm{Im}\rho)^3}. \tag{28.35}$$

We will not derive this formula, only argue about the factor $a_0^4 T_3$: here T_3 is the 3-brane tension, so the energy should be proportional to it, and a_0 is the warp factor, so for a 4-dimensional world-volume, we should indeed have this factor (the volume of the anti-D3-brane worldvolume is a_0^4 times the volume of the Minkowski coordinates).

Then, because of the a_0^4 factor, the anti-D3-brane will run off to the minimum of a_0, $a_{\min}$, the end of the KS throat (or cigar). For more general validity, we write

$$\delta V = \frac{D}{(\mathrm{Im}\rho)^3}. \tag{28.36}$$

This potential is to be added to the potential coming from the supersymmetric case, obtaining

$$\begin{aligned} V &= \frac{1}{(2\sigma)^3}\left[\frac{4\sigma^2}{3}\frac{9}{4\sigma^2}\left(W_0 + Ae^{-a\sigma}\left(1+\frac{2a\sigma}{3}\right)\right)^2 - 3(W_0+Ae^{-a\sigma})^2\right] + \frac{D}{\sigma^3} \\ &= \frac{Ae^{-a\sigma}}{(2\sigma)^3}\left[6\left(1+\frac{2a\sigma}{3}\right)W_0 + 3Ae^{-a\sigma}\left(1+\frac{4a\sigma}{3}+\frac{4a^2\sigma^2}{9}\right) - 6W_0 - 3Ae^{-a\sigma}\right] + \frac{D}{\sigma^3} \\ &= \frac{aAe^{-a\sigma}}{2\sigma^2}\left[W_0 + Ae^{-a\sigma}\left(1+\frac{a\sigma}{3}\right)\right] + \frac{D}{\sigma^3}. \end{aligned} \tag{28.37}$$

The shape of the resulting potential, with the parameters chosen as explained above, is: the potential drops from infinity at zero, down to a positive value (de Sitter vacuum), then to a maximum, and then runs off to zero at infinity. One can in fact fine tune the coefficient D of the susy breaking term to obtain a vacuum value V_0 very close to zero, consistent with what we see now (very small cosmological constant).

The fine tuning of D is not as extreme as in the cosmological constant, but is fine nevertheless. Note that D is the parameter responsible for supersymmetry breaking, which lifts a would-be AdS background to a dS background.

Important Concepts to Remember

- In supergravity, supersymmetric vacua are AdS vacua.
- With a volume modulus ρ for KK compactification and W independent of ρ, we obtain a positive definite potential $V \geq 0$, with a Minkowski vacuum.
- But there are nonperturbative contributions to $W(\rho)$ from Euclidean D3-brane instantons and gluino condensation on D7-branes wrapping 4-cycles, giving generically $W = Ae^{ia\rho}$, with $a > 0$, besides the constant W_0 perturbative term.
- The moduli are stabilized by integer fluxes on the compact space.
- The solution is a compactified version of the Klebanov–Strassler geometry, where the "cigar" geometry in y has a minimum volume at one sides, and is cut off and glued onto a half of CY_3 at the other.
- The supersymmetric minimum is AdS, but by adding antibranes sitting at the tip of the cigar, with $V = D/(\mathrm{Im}\rho)^3$, we get a de Sitter minimum.

Further reading: The original KKLT paper is [49].

Exercises

(1) Consider a superpotential

$$W = (A\rho + B)e^{ia\rho} + C \ , \tag{28.38}$$

where $A \neq 0$, $a \in \mathbb{R}_+$. Find the conditions for a supersymmetric minimum.
(2) Add an anti-D3-brane to the previous exercise. Do we obtain a dS vacuum?
(3) Consider the superpotential

$$W = A\eta(\rho)^{-6}e^{\frac{3i\tau}{4b}} + B \ , \tag{28.39}$$

where $\eta(x)$ is the Dedekind eta function. Is the volume stabilized? How about the dilaton?
(4) Find a way to obtain a dS vacuum for the superpotential at exercise 3 (by breaking supersymmetry in a controllable way).

Chapter 29
The KKLMMT Scenario for Inflation and Generalizations

In the previous chapter, we have shown how KKLT have obtained a de Sitter background in string theory, with a geometry made from branes, cut off in the UV and smoothly joined onto a CY_3 half of space, and an antibrane at the tip (minimum of the KS throat). In this chapter, we will present the KKLMMT scenario for string inflation based on it, of the type of brane-antibrane inflation already analyzed before in the book.

Since the potential energy found by KKLT depended on the volume modulus ρ, it would have been a natural assumption to believe that we could construct an inflationary scenario for a potential for ρ. Or perhaps by letting the antibrane slide towards the tip of the KS geometry r_0, with its position being the inflaton. But in fact, the inflaton is neither: the volume modulus ρ becomes stabilized at the de Sitter minimum, as we saw, and the antibrane has too steep a potential, leading to the eta problem ($\eta \sim 1$). In fact, in order to obtain inflation, we must introduce a sliding D3-brane in the KS throat geometry, and then its position r_1, or rather the brane-antibrane separation $(r_1 - r_0)$, acts as the inflaton, just like in usual brane-antibrane inflation.

Note that, since the KS geometry itself is made from D3-branes, if we would put only the sliding D3-brane in the geometry, we would have no potential, since the solution would be supersymmetric (we can always add parallel branes without breaking extra supersymmetry. But because there is an anti-brane, breaking susy, we get a potential: the anti-D3 modifies the background (there is a back-reaction), and the D3-brane feels this modification.

This would give inflation as it stands, but we will see that we need to fix the other moduli, and then this generically will spoil the inflationary potential, and we would get $\eta \sim 1$. So in fact, as emphasized in the KKLMMT paper, the scenario doesn't really describe inflation in string theory, but rather the opposite: that generically, we *don't* obtain inflation, if we want to stabilize all the moduli.

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_29

29.1 Simplified Model for KKLMMT

We will start with the analysis of a simplified model, where we replaced the somewhat complicated geometry of the KS throat with a simpler solution, $AdS_5 \times X^5$, coming from N D3-branes. But then we must *phenomenologically* cut off the radial r direction of AdS_5 in the IR at r_0 and the UV at $r_{\max}$, in order to make the space compact, thus creating a Randall–Sundrum (RS) like construction (the RS 2-brane model has an IR and a UV brane cutting off a slice of AdS space). The cut-offs r_0 and $r_{\max}$ can be chosen independently of R (as we will see later in the KS construction). The compact space X^5 is a generalization of the pure D3-brane case, that gives S^5. We can have X^5 be a Sasaki–Einstein space in general, for instance a space obtained by dividing (making an equivalence under) a discrete group like $\mathbb{Z}_2$.

The metric of the space is

$$ds^2 = \frac{r^2}{R^2}(-dt^2 + d\vec{x}^2) + \frac{R^2}{r^2}dr^2 + ds^2_{X^5}\ , \tag{29.1}$$

where the radius is defined by an integer N (the flux through AdS, or the number of D3-branes generating it), as

$$R^4 = 4\pi a g_s N \alpha'^2. \tag{29.2}$$

Here a is a phenomenological parameter that allows us to cover several cases, including the KS case. In the pure D3-brane case, $a = 1$, but for KS we will see that $a \neq 1$.

The sliding D3-brane is put at a large distance from the tip where we have the anti-D3-brane, i.e.,

$$r_0 \ll r_1. \tag{29.3}$$

Calculation of the Potential

In order to calculate the potential, we first rewrite the metric in the way that it was originally described, as coming from N D3-branes, namely in terms of a harmonic function $H_3(r)$,

$$\begin{aligned} ds^2 &= H_3^{-1/2}(r)(-dt^2 + d\vec{x}^2) + H_3^{1/2}(r)\left(dr^2 + \frac{r^2}{R^2}\tilde{g}_{ab}dy^a dy^b\right) \\ ds^2_{X^5} &= \tilde{g}_{ab}dy^a dy^b \\ H_3(r) &= \frac{R^4}{r^4}. \end{aligned} \tag{29.4}$$

Note that the harmonic function is a harmonic function of the transverse space,

$$ds^2_6 = dr^2 + \frac{r^2}{R^2}\tilde{g}_{ab}dy^a dy^b\ , \tag{29.5}$$

which for $X^5 = S^5$ is simply flat 6 dimensional space, and the harmonic function is the usual one, $1 + R^4/r^4$, with the 1 dropped because we are in the near-horizon ($r \to 0$) limit.

Moreover, we have a 4-form potential coupling to the D3-brane of value

$$\begin{aligned}(C_4)_{tx^1x^2x^3x^4} &= \frac{r^4}{R^4} \equiv H_3^{-1}(r) \Rightarrow \\ (F_5)_{rtx^1x^2x^3x^4} &= \partial_r H_3^{-1}(r) = \frac{4r^3}{R^4}.\end{aligned} \tag{29.6}$$

As we said, in order to calculate the potential, we should calculate the back-reaction effect of the anti-D3-brane on the background, and then calculate the action of the sliding D3-brane in the modified background. But since the force exerted by the D3-brane on the anti-D3-brane equals the force exerted by the anti-D3-brane on the D3-brane, we can calculate the potential due to the D3 and anti-D3 by reversely considering how the sliding D3-brane modifies the background, and calculating the action of the anti-D3-brane in the modified background.

This is easier to do, since the harmonic function of parallel branes with no anti-branes (a supersymmetric configuration) is simply the sum of the harmonic functions. So we must make in the above solution the substitution

$$H_3(r) \to H_3'(r) = \frac{R^4}{r^4} + \delta h(r)\ , \tag{29.7}$$

where $\delta h(r)$ is again a harmonic function of the transverse space. The coefficient of $\delta h(r)$ is one instead of N, and centered at position r_1 instead of $r = 0$, i.e.,

$$\begin{aligned}\Delta_6 \delta h(r) &= C\delta^{(6)}(\vec{r} - \vec{r}_1) \Rightarrow \\ \delta h(r) &= \frac{R^4}{N} \frac{1}{|\vec{r} - \vec{r}_1|^4} \simeq \frac{R^4}{N} \frac{1}{r_1^4}\ ,\end{aligned} \tag{29.8}$$

where in the last equation we have used that $r \ll r_1$, since $r = r_0$ is the position of the anti-D3-brane (where we will calculate the action). In conclusion, we have

$$H_3'(r) = R^4 \left(\frac{1}{r^4} + \frac{1}{N} \frac{1}{r_1^4} \right). \tag{29.9}$$

In this modified background, we calculate the potential by evaluation the action of the anti-D3-brane (the DBI plus the WZ term) at the position $r = r_0$,

$$\begin{aligned}S_{\text{DBI+WZ,D3}/\overline{\text{D3}}-\text{brane}}(r = r_0) = &-T_3 \int \sqrt{-g_{(4)}} d^4x H_3^{-\frac{p+1}{4}}(r_0) \sqrt{1 + H_3(r_0) \partial_\mu r_0 \partial^\mu r_0} \\ &\pm T_3 \int (C_4)_{tx^1x^2x^3x^4} dt dx^1 dx^2 dx^3 dx^4.\end{aligned} \tag{29.10}$$

The plus (upper) sign is for the D3-brane, when the leading term in DBI cancels against the leading term in the WZ term, and ignoring $\delta h(r)$, we have only the kinetic term,

$$S \simeq -T_3 \int d^4x \frac{r_0^4}{R^4} \sqrt{1 + \frac{R^4}{r_0^4} \partial_\mu r_0 \partial^\mu r_0} \pm T_3 \int (C_4)_{t x^1 x^2 x^3 x^4} dt dx^1 dx^2 dx^3 dx^4$$
$$\simeq -\frac{1}{2} \int d^4x T_3 \partial_\mu r_0 \partial^\mu r_0. \tag{29.11}$$

We have used the fact that the 4 dimensional metric is flat, and we see that we obtain the canonical kinetic term, modulo a trivial rescaling, such that the canonical scalar is

$$\phi = \sqrt{T_3} r_0. \tag{29.12}$$

As we have already shown, the 3-brane tension is

$$T_3 = \frac{1}{(2\pi)^3 g_s \alpha'^2}. \tag{29.13}$$

Then, by introducing the modified harmonic function $H_3'(r)$ and considering the anti-D3-brane case at $r = r_0$, we find the potential

$$V(r_0, r_1) = 2T_3 \frac{r_0^4}{R^4} \left(1 - \frac{1}{N} \frac{r_0^4}{r_1^4}\right) , \tag{29.14}$$

which we leave as an exercise to prove.

While this potential is in principle a function of both r_0 and r_1, only r_1 can vary, and in the physical KS case r_0 is fixed.

Then, as a function of the inflaton ϕ, we obtain the slow-roll parameters

$$\epsilon \equiv \frac{M_{\rm Pl}^2}{2} \left(\frac{V'}{V}\right)^2 = \left(\frac{M_{\rm Pl}^2}{T_3 r_1^2}\right) \frac{1}{2N} \left(\frac{4 r_0^4}{r_1^4}\right)^2 \ll 1$$
$$\eta \equiv M_{\rm Pl}^2 \left(\frac{V''}{V}\right) = -\left(\frac{12 M_{\rm Pl}^2}{T_3 r_1^2}\right) \frac{1}{N} \left(\frac{r_0}{r_1}\right)^4 \ll 1. \tag{29.15}$$

We have obtained $\epsilon, \eta \ll 1$ since all the 3 factors in each are $\ll 1$, including

$$\frac{M_{\rm Pl}^2}{T_3 r_1^2} = \frac{(2\pi)^3 M_{\rm Pl}^2 g_s \alpha'^2}{r_1^2} \ll 1 , \tag{29.16}$$

where the ratio is much smaller than one, since if the distance r_1 (brane-antibrane distance) is not much larger than the Planck and string scales, we would need to consider quantum string corrections.

29.2 String Theory Example for KKLMMT Construction

We now move on to the standard example in string theory, where instead of the $AdS_5 \times S^5$ space we have the KS geometry. This is an example in type IIB string theory, with G-flux, i.e., with flux for

$$G_3 = F_3 - \tau H_3 \ , \tag{29.17}$$

on 3-cycles in the internal space. As described in the previous chapter, the total KS space is written as

$$ds^2 = e^{2A(y)}\eta_{\mu\nu}dx^\mu dx^\nu + e^{-2A(y)}g_{mn}(y)dy^m dy^n. \tag{29.18}$$

At the tip of KS throat, there is an S^3 of finite size, leading to the fact that the warp factor $a = e^A$ has a minimum. Then, at a certain $y = y_{\max}$, we cut off the KS space, and glue half of a CY_3 space, specifically a fibration over the 5 dimensional $T^{1,1}$ space (I will not explain what it is), as in Fig. 29.1.

There are integer G-fluxes defined on 3-cycles A and B,

$$\frac{1}{(2\pi)^2\alpha'}\int_A F_3 = M; \quad \frac{1}{(2\pi)^2\alpha'}\int_B H_3 = -K \ , \tag{29.19}$$

where A is the S^3 at the tip of the KS geometry, and B is a dual 3-cycle. We can choose $M, K \gg 1$. The KS geometry is not known analytically over the whole range of r, only near the tip, and far away from the tip. Far away from the tip, we have the metric

$$ds^2 = H^{-1/2}(r)\eta_{\mu\nu}dx^\mu dx^\nu + H^{+1/2}(r)(dr^2 + r^2 ds^2_{T^{1,1}}) \ , \tag{29.20}$$

and the harmonic function is a log-corrected $AdS_5 \times S^5$ with $a \neq 1$,

$$H(r) = \frac{27\pi}{4r^4}\alpha'^2 g_s M\left[K + g_s M\left(\frac{3}{8\pi} + \frac{3}{2\pi}\ln\frac{r}{r_{\max}}\right)\right]. \tag{29.21}$$

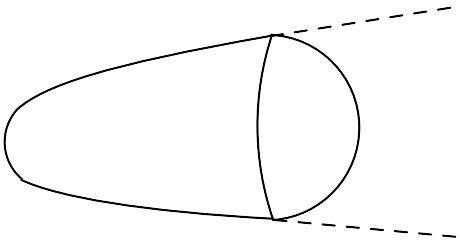

Fig. 29.1 KKLMMT construction: take the KS "cigar geometry", or semi-infinite almost cone with a smooth tip, and cut it off at some position, and glue half a CY space there, to make it compact

As we see, the leading term (without the log correction) is an AdS space defined by a radius of the type already used,

$$R^4 = \frac{27}{16} 4\pi g_s N \alpha'^2 \ , \tag{29.22}$$

so with $a = 27/16$, and where $N \equiv MK$ is the *product* of the two fluxes. The presence of $T^{1,1}$ in the metric means that we can *smoothly* join a half of a CY_3 space which is a fibration over $T^{1,1}$ onto the KS throat cut off at some finite r. It also means that the volume modulus of the CY will then be proportional to the volume modulus of the whole resulting compact space.

As explained in the previous chapter, the warp factor has a minimum, but the minimum is defined by the *ratio* of fluxes, as

$$\left(\frac{r_0}{R}\right)^4 = a_0^4 \equiv a_{\min}^4 = e^{-\frac{8\pi K}{3g_s M}} . \tag{29.23}$$

This ends the description of the "model for inflation". We leave as an exercise to calculate the inflaton potential in this physical KS case.

29.3 Volume Stabilization and Problems

We now turn to the problems: when we try to stabilize the moduli, we generically obtain an $\eta \sim \mathcal{O}(1)$. Consider the complex volume modulus ρ, rescaled as

$$\rho_{\text{KKLMMT}} = -i\rho_{\text{KKLT}} \ , \tag{29.24}$$

and the complexified D3-brane position ϕ that becomes the inflaton. Then the Kähler potential for ρ is, as we already explained, simply from KK compactification, is $K = -3\log(\rho + \bar{\rho})$. Adding the D3-brane moduli, as we also explained, is done by adding a Kähler potential $\kappa(\phi, \bar{\phi})$ inside the log,

$$K(\rho, \bar{\rho}, \phi, \bar{\phi}) = -\log(\rho + \bar{\rho} - \kappa(\phi, \bar{\phi})) \ , \tag{29.25}$$

where $\kappa(\phi, \bar{\phi})$ starts with a canonical term,

$$\kappa(\phi, \bar{\phi}) = \bar{\phi}\phi + \cdots \tag{29.26}$$

A partial explanation for the above structure is that the moduli space (the leading kinetic term $g_{ij}\partial\phi^i\partial\phi^j$) of D-branes has metric

$$ds^2 = \frac{3}{2r^2}\left(dr^2 + \left(d\chi + \frac{i}{2}\partial_j\kappa d\phi^i - \frac{i}{2}\partial_{\bar{j}}\kappa d\bar{\phi}^{\bar{j}}\right)^2\right) + \frac{3}{r}\partial_i\partial_{\bar{j}}\kappa d\phi^i d\bar{\phi}^{\bar{j}} . \tag{29.27}$$

Here $r \sim e^{4u}$ is proportional to the volume of the CY space (and thus of the whole compact space). One would be tempted to work with the complex modulus $r + i\chi$, but that is not generated by a Kähler potential. Instead, one must choose the complex parameter ρ, whose imaginary part is the axion (as we said in the previous chapter), while the real part is defined by

$$2r = \rho + \bar{\rho} - \kappa(\phi, \bar{\phi}) \simeq \rho + \bar{\rho} - \phi\bar{\phi} + \cdots . \tag{29.28}$$

If, as we said, the perturbative superpotential equals the tree level contribution, which is constant,

$$W = W_0 , \tag{29.29}$$

then, since (a labels ϕ^i and ρ)

$$g^{a\bar{b}} \partial_a K \partial_{\bar{b}} K = 3 , \tag{29.30}$$

we obtain $V = 0$, so indeed ρ and ϕ are *moduli*.

If however, $W = W(\phi^i)$ only, and stabilizes ϕ^i, then *in a supersymmetric vacuum*, i.e., with $\partial_\phi W = 0$, the potential will be independent of ρ, so ρ is still a modulus.

Note that the Kähler potential K that we wrote does not have good transformation properties in the limit that K is given by κ only, namely at large enough (and stabilized) volume, so that the D-brane will move approximately in flat space, i.e., will have a canonical kinetic term. That however can be fixed by assigning the modified Kähler transformation for κ,

$$\begin{aligned} &\kappa(\phi, \bar{\phi}) \to \kappa + f(\phi) + \overline{f(\phi)} \\ &\rho \to \rho + f, \quad \bar{\rho} \to \bar{\rho} + \bar{f}. \end{aligned} \tag{29.31}$$

Finally, the volume modulus ρ can be stabilized by the nonperturbative contribution of the KKLT model (see the previous chapter),

$$W = W_0 + A e^{-a\rho_{\text{KKLMMT}}} = W_0 + A e^{ia\rho_{\text{KKLT}}}. \tag{29.32}$$

But this stabilization has the generic problem that for volume $\propto e^{4u} \propto r$ stabilized at large r, the potential will be generically of the type

$$V(r, \phi) \sim \frac{X(\rho)}{r^\alpha} = \frac{X(\rho)}{(\rho - \bar{\phi}\phi/2 + \cdots)^\alpha}. \tag{29.33}$$

Then if ρ is stabilized at $\rho \simeq \rho_0$, by expanding the above potential around $\phi = 0$ and considering zero axion (imaginary part of ρ is zero, so ρ is real) we obtain

$$V \simeq V_0 \left(1 + \alpha \frac{\phi\bar{\phi}}{2\rho_0} + \cdots \right) , \tag{29.34}$$

where $\alpha \sim \mathcal{O}(1)$, and also $\rho_0 \sim \mathcal{O}(1)$ in Planck units (the volume modulus stabilized close to the Planck scale, as is natural), then $\eta \sim \mathcal{O}(1)$.

Of course, if instead of $X(\rho)$ we have a more general $X(\rho, \phi)$, then we will get an extra contribution

$$V \simeq V_0 \left(1 + \alpha \frac{\phi\bar{\phi}}{2\rho_0} + \frac{\partial_\phi \partial_{\bar{\phi}} X(\rho_0, 0)}{X(\rho_0, 0)} \phi\bar{\phi} + \cdots\right) , \tag{29.35}$$

and by fine-tuning we could arrange for the two terms to (almost) cancel, so that we get $\eta \ll 1$. But it means that this doesn't happen generically.

29.4 KKLMMT Generalization with Magnetic Flux

Since volume stabilization is such an important issue, we must look for generalizations. One such generalization is in the paper by Abe, Higaki and Kobayashi [51], which shows that we can have not just the $Ae^{i\rho_{\rm KKLT}}$ superpotential contribution with $a > 0$, but also one with

$$Be^{-ia\rho} , \tag{29.36}$$

again with $a > 0$, and with

$$B = Ce^{-m_9 c\langle S\rangle}. \tag{29.37}$$

Here the dilaton modulus is S,

$$2\pi S = e^{-\phi} - ic_0 , \tag{29.38}$$

where c_0 is the axion,

$$c = \frac{8\pi^2}{N_9}. \tag{29.39}$$

This superpotential comes from gluino condensation on magnetized D9-branes added to the KKLT construction, such that the coupling constant of the theory on the D9-branes is given by

$$\frac{1}{g_{D9}^2} = |m_9 {\rm Re}S - w_9 {\rm ReT}| , \tag{29.40}$$

and $T = -i\rho_{\rm KKLT} = \rho_{\rm KKLMMT}$.

Moreover, a term of this type in the superpotential would be required by imposing the condition of *T-duality invariance* (as well as the condition for "modular invariance" about which we will not say anything). Indeed, we saw that T-duality, relating small compact volumes with large compact volumes, is a symmetry of string theory, so the potential, that blows up at very small volumes, should also blow up at large

ones. In fact, it was argued that the superpotential arising from gaugino condensation from compactification on tori is

$$W(T, S) \sim \eta(iT)^{-6} e^{-\frac{3S}{8\pi b}} , \tag{29.41}$$

where $\eta(x)$ is the Dedekind eta function. At large volume, Re $T \to \infty$, the superpotential behaves as

$$W \propto e^{\pi T/2} , \tag{29.42}$$

with a proportionality coefficient of the order of

$$\sim e^{-\frac{3S}{8\pi b}} , \tag{29.43}$$

and $b \sim \mathcal{O}(1)$ is an RG flow factor.

We should mention however that there is some controversy about this calculation. Despite T-duality invariance being needed, it was argued that it is unphysical (from the point of view of the string coupling dependence in string theory) to have the potential blowing up as the volume blows up.

29.5 Type IIA Model Corresponding to KKLMMT

Until now we have worked in type IIB string theory, but an alternative model in type IIA string theory has been proposed.

The background that replaces KS in this case is the doubly Wick rotated nonextremal D4-brane background considered before, when discussing brane inflation, namely

$$\begin{aligned} ds^2 &= H_4^{-1/2}(r)(-dt^2 + d\vec{x}_3^2 + f(r)dx_6^2) + H_4^{+1/2}(r)\left(\frac{dr^2}{f(r)} + r^2 d\Omega_4^2\right) \\ C_{(5)} &= \frac{1}{g_s} H_4^{-1}(r) dt \wedge dx^1 \wedge \cdots \wedge dx^4 \\ f(r) &= 1 - \left(\frac{r_H}{r}\right)^3 \\ H_4(r) &= 1 + \alpha_4 \left(\frac{r_4}{r}\right)^3 . \end{aligned} \tag{29.44}$$

The geometry is now:

- a cigar geometry for r over S^4, with a physical cut-off at $r_{\min} = r_H$, and an explicit cut-off at $r_{\max}$.
- a cigar geometry for r over x_6, again cut off at $r_{\max}$, and at r_H.

Then, at $r_{\max}$, we can glue a half of a CY_3 space W over the space (r, x_6, S^4), and finally find a construction that is T-dual to KKLT.

Model 1

We can construct a model by adding a sliding D4-brane in the above construction. Even though it looks different, it is actually the same construction we had before, since the fact that the background is nonextremal means that we have anti-D4-branes in it (instead of the probe anti-D4-brane situated at the tip of the geometry).

Model 2

We can construct an alternative model by putting $r_H = 0$, i.e., considering a supersymmetric background, without any anti-D4-branes in it, but then adding a sliding anti-D4-brane probe to break supersymmetry and generate a potential.

It is left as an exercise to show that the potentials we obtain in the two models, by calculating as before the DBI+WZ action for the probe in the background, are

$$V_1(\phi)|_{\text{plateau}} \simeq \frac{T_4(2\pi R)}{g_s}\left(\frac{\phi_H}{\phi_4}\right)^3\left[N - \frac{1}{2\alpha_4}\left(1 + \frac{\phi_4^3}{4\phi^3} - \frac{\phi^3}{\alpha_4\phi_4^3}\right)\right]$$
$$V_2(\phi) = \frac{2T_4(2\pi R)}{g_s}\frac{1}{1+\frac{\phi_4^3}{\phi^3}}. \tag{29.45}$$

Important Concepts to Remember

- The KKLMMT scenario is a modification of the KKLT scenario, where the volume modulus (in the presence of the anti-D3-brane at the tip) is stabilized at its de Sitter minimum, and the inflaton is a sliding D3-brane, or rather the D3-anti-D3 separation, like in usual brane-antibrane inflation.
- A simplified model has an $AdS_5 \times X^5$ space, cut off at r_0 and $r_{\max}$ (chosen independently of R, the radius of the geometry), coming from N D3-branes.
- To calculate the potential, instead of calculating the backreaction of the anti-D3 on the geometry, and then the potential for the sliding D3, we calculate how the sliding D3 modifies the background, and then the potential for the fixed anti-D3 in the modified geometry.
- The resulting potential has a plateau followed by a drop, $V(\phi) \sim V_0(1 - \alpha/\phi^4)$, and has $\epsilon, \eta \ll 1$.
- In the physical string theory case, with the KS geometry, still r_0 and $r_{\max}$ are chosen independently, as r_0 depends on a ratio of fluxes K/M, whereas the radius R on the product $N = MK$, and $r_{\max}$ is arbitrary.
- When we try to stabilize the volume modulus ρ through a nonperturbative superpotential like the KKLT one, generically we obtain a $\mathcal{O}(1)$ contribution to η, i.e., a quadratic contribution to $V(\phi)$ of order one in Planck units.

- We can also generate a $W = Be^{-ia\rho}$ (opposite sign in the exponent) for the KKLT scenario by adding magnetized D9-branes. Moreover, such a term would be required for T-duality invariance. The resulting potential has a minimum (it blows up in the IR and the UV).
- A type IIA version of KKLT can be obtained by replacing the KS geometry with the doubly Wick rotated nonextremal D4-brane geometry,
- One model has a sliding D4-brane in the nonextremal geometry, another has a sliding anti-D4-brane in the extremal geometry.

Further reading: The original KKLMMT paper is [50]. The KKLT model with magnetic flux is [51], and the type IIA model is [35].

Exercises

(1) Prove that the KKLMMT potential for the $AdS_5 \times S^5$ background is

$$V(r_0, r_1) = \frac{2T_3 r_0^4}{R^4}\left(1 - \frac{1}{N}\frac{r_0^4}{r_1^4}\right). \tag{29.46}$$

(2) Repeat the exercise for the RS background, to find its $V(r_0, r_1)$.
(3) Check that for the case T-dual to KKLMMT, we have

$$\begin{aligned} V_1(\phi)|_{\text{plateau}} &\simeq \frac{T4(2\pi R)}{g_s}\left(\frac{\phi_H}{\phi_4}\right)^3\left[N - \frac{1}{2\alpha}\left(1 + \frac{\phi_H^3}{4\phi^3} - \frac{\phi^3}{\alpha_4\phi_4^3}\right)\right] \\ V_2(\phi) &= \frac{2T_4(2\pi R)}{g_s}\frac{1}{1 + \frac{\phi_4^3}{\phi^3}}. \end{aligned} \tag{29.47}$$

(4) Consider the modification of the KKLMMT model with superpotential

$$W(T, S) \sim \eta(iT)^{-6}e^{-\frac{3S}{8\pi b}}\ , \tag{29.48}$$

and calculate the scalar potential, and check whether all scalars are stabilized.

Chapter 30
The Ekpyrotic Scenario

In this chapter we will describe the ekpyrotic scenario, which is an alternative to inflation, inspired from string theory, defined in a 2001 paper by Khoury, Ovrut, Steinhardt and Turok. The general idea is that we have two "end of the world" branes that move towards each other in a fifth direction, and their collision represents the Big Bang.

In fact, the original idea of the ekpyrosis (coming from the ancient Greek word for the end of the Universe in a conflagration) came from a string theory construction (shown in the original paper, and explained in more detail later): In heterotic M-theory, the strong coupling version of the heterotic string, the "eleventh" direction, of radius $R = g_s^{2/3}$, has two "end of the world" M9-branes at its ends. In fact, the eleventh direction is an interval $S^1/\mathbb{Z}_2$, obtained by identifying ("orbifolding") a circle under reflection, which means that the M9-branes at the endpoint sit on the orbifold fixed points at the end of the interval. Each M9-brane has an E_8 group on them, for a total $E_8 \times E_8$ gauge group. But in this construction, there is also an M5-brane in the bulk, moving towards one of the M9-branes, and finally colliding with it. However, we will see that this mechanism is not favoured in modern ekpyrotic constructions, and instead we must have only two "end of the world" branes moving towards each other and colliding.

So it would seem like it has some similarities with brane-antibrane inflation, but that is only superficial. In the brane-antibrane inflation scenario the Universe is always expanding, as the separation decreases, and the collision corresponds to reheating; moreover, contrary to what will be seen here, the potential was positive, leading to an inflating Universe, with the inflaton being the separation.

By contrast, now almost everything relevant happens *before* the Big Bang, in a *contracting phase* with *negative* potential. The primordial fluctuations are created not by quantum fluctuations in some scalar during the expanding phase, but rather by fluctuations on the branes, *before the collision*.

We now solve in a different way the smoothness problem (the fact that the Universe is uniform and isotropic on the largest scales) and the related horizon problem, which means that when extrapolating in the past the current cosmology we get that there are many patches in the current sky that were causally disconnected (72 patches that were

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_30

disconnected on the surface of last scattering), and yet they seem to be correlated. These problems are now solved because the branes colliding are held together by tension as a single object, and the collision happens at every point on the branes in the same time. So the connectedness of various regions is due to the tension, or said another way, due to the fact that the Universe is much older than the Big Bang, so there was some causal connection of the various points *before the Big Bang*. On the other hand, in inflation these puzzles are solved by the exponential expansion due to the inflaton, *after the Big Bang*.

Like in the case of inflation, we can have an *effective field theory description* for most of the evolution, in which the fifth dimension is integrated over (somewhat similar to what happens in dimensional reduction), and the interbrane separation becomes a scalar field in 4 dimensions, $\phi(t, \vec{x})$. We can explain the early stages of the ekpyrotic scenario entirely in terms of evolution of the scalar ϕ coupled to gravity, just like in inflation. However, the later stages must take into consideration the actual higher dimensional set-up.

30.1 The Ekpyrotic Phase

The essential tool of the ekpyrotic model is the existence of a *contracting* ekpyrotic phase. It is due to a potential that has the form

$$V = -V_0 e^{-c\phi} , \tag{30.1}$$

which can be obtained as the potential for the interaction of the bulk brane with the boundary brane in the original heterotic M-theory construction sketched above. However, it was realized that we only need to have the above potential, not necessarily the construction with the bulk brane (though without the bulk brane it becomes unclear how to justify the above potential).

Note that the above potential, *if considered with a positive sign*, gave an exact solution for inflation (as we saw in Chap. 9, of part I). In fact, we can use the same algebra to solve it also now.

The KG equation for the scalars (we consider several scalars for generality) and the two Friedmann equations (the second being the dependent one) are

$$\begin{aligned} &\ddot{\phi} + 3H\dot{\phi} + V_{i,\phi_i} = 0 \\ &H^2 = \frac{1}{3}\left[\sum_i \frac{1}{2}\dot{\phi}_i^2 + \sum_i V_i(\phi_i)\right] \\ &\dot{H} = -\frac{1}{2}\sum_i \dot{\phi}_i^2 . \end{aligned} \tag{30.2}$$

Here the potential is $V = \sum_i V_i(\phi_i)$, where

$$V_i(\phi_i) = -V_i e^{-c_i \phi_i} \ , \tag{30.3}$$

and we consider $c_i \gg 1$ for all i. Then we can show (which we leave as a simple exercise) that there is a scaling solution,

$$\begin{aligned} a(t) &= (-t)^p \\ \phi_i &= \frac{2}{c_i} \log \left(-t \sqrt{\frac{c_i^2 V_i}{2}} \right) \\ p &= \sum_i \frac{2}{c_i^2} \end{aligned} \tag{30.4}$$

Here $t < 0$, so the evolution is from $t = -\infty$ to $t = 0$, i.e., towards a Big Crunch, after which we have the Big Bang, and our current cosmology. Since we must have $c_i \gg 1$ for all i, it follows that $p \ll 1$, so the time evolution is very slow ($a(t) \simeq$ const.). Note that by substituting ϕ_i in the potential, we obtain the time evolution

$$V_i(\phi_i(t)) = -\frac{2}{c_i^2 t^2}. \tag{30.5}$$

On the other hand, comparing $a(t)$ with the evolution due to a single matter component with equation of state $P = w\rho$, which is

$$a(t) \propto t^{\frac{2}{3(1+w)}} \ , \tag{30.6}$$

we obtain

$$w = \frac{2}{3p} - 1 \gg 1. \tag{30.7}$$

Thus we need matter with $w \gg 1$ in order to obtain the ekpyrotic phase.

Note that for a single i,

$$w = \frac{2}{3p} - 1 = \frac{c^2}{3} - 1 \ , \tag{30.8}$$

so the condition $w > 1$ (certainly needed if $w \gg 1$, but also will be needed later) implies

$$c > \sqrt{6}. \tag{30.9}$$

Solution to the Flatness Problem

As we said, for $p \ll 1$, we have $a(t) \simeq$ constant, but on the other hand we obtain

$$H = \frac{\dot{a}}{a} \propto \frac{1}{t} \ , \tag{30.10}$$

which means that we can solve the flatness problem (the fact that the Universe is approximately flat near the Big Bang, $\Omega \simeq 1$, independent of what we have now) in a way similar to that done during inflation: In inflation it was solved by having aH grow by e^{60} (60 e-folds), though that was done by H being constant and a increasing by e^{60}. We can solve it now again by having aH increase by 60 e-folds, but rather by having a constant, and instead H increasing by e^{60}, which in a contracting phase means that

$$|t_{\rm beg}| \geq e^{60}|t_{\rm end}|. \tag{30.11}$$

Note that the end time of the ekpyrotic phase can be close enough to the Planck scale, so that the condition on $|t_{\rm beg}|$ is not necessarily very stringent.

We mentioned that $w > 1$ is needed, and in fact in order to avoid chaotic behaviour, the Belinsky–Khalatnikov–Lifshitz (BKL) singularity, an instability towards an anisotropic, homogenous, chaotically oscillating solution (the anisotropies oscillate), we need $w > 1$.

During the ekpyrotic phase, the Universe contracts, but the energy density of a component with equation of state $P = w\rho$ behaves as

$$\rho \propto a^{-3(1+w)} , \tag{30.12}$$

which means that as t evolves, the highest w will dominate (whereas in the usual expanding phase, the lowest w dominates asymptotically: the cosmological constant with $w = -1$ gets to dominate eventually). That justifies the choice of $w \gg 1$ for the ekpyrotic phase: it is the last w to dominate during the contraction.

30.2 Kinetic Phase

Further, as one approaches even more the Big Crunch (and thus the Planck scale), eventually the potential (which is defined by some energy scale) must become irrelevant (the potential cannot increase without bound, but the kinetic energy will increase), so we have

$$S = -\frac{1}{2}\int d^4x\sqrt{-g}(\partial_\mu\phi)^2 , \tag{30.13}$$

leading to the energy-momentum tensor

$$T_{\mu\nu}(\phi) = \partial_\mu\phi\partial_\nu\phi - \frac{1}{2}g_{\mu\nu}(\partial_\rho\phi)^2. \tag{30.14}$$

Considering a uniform scalar, so $\phi = \phi(t)$, we obtain

$$T_{00} = T_{ii} = \frac{1}{2}\dot{\phi}^2 \Rightarrow P = \rho \Rightarrow w = 1. \tag{30.15}$$

This is called (as we explained in part I of the book) stiff matter. The KG and Friedmann equations for this case are

$$\begin{aligned} &\ddot{\phi} + 3H\dot{\phi} = 0 \\ &3H^2 = \frac{1}{2}\dot{\phi}^2 = -\dot{H}. \end{aligned} \tag{30.16}$$

From the second equation, we obtain

$$a(t) \propto e^{\frac{\phi(t)}{\sqrt{6}}}, \tag{30.17}$$

whereas from $3H^2 = -3\dot{H}$ we get

$$a(t) = a_0(-t)^{1/3}. \tag{30.18}$$

Comparing the two, we find

$$\phi(t) = \sqrt{\frac{2}{3}}\ln(-t) + \phi_0. \tag{30.19}$$

Finally, we find that indeed, as for stiff matter

$$\rho = \frac{\dot{\phi}^2}{2} \propto a^{-6}. \tag{30.20}$$

At the end of the kinetic phase, we should reach the Planck scale. That means that before that, the fifth dimension will become important in the analysis. In turn, that means that it is important to remember that ϕ represents the distance between the colliding branes in the fifth dimension, which is the length of a line segment $S^1/\mathbb{Z}_2$.

In KK dimensional reduction from dimension D down to dimension d, as we saw in Chap. 13, the nonlinear ansatz for gravity (that gives the correct Einstein action in the lower dimension) is

$$ds_D^2 = \Delta^{-\frac{1}{d-2}} ds_d^2 + g_{mn} dx^m dx^n, \tag{30.21}$$

where $\Delta = \det g_{mn}$. When reducing from $D = 5$ down to $d = 4$, since we have a single extra dimension, $g_{55} = R^2$, so the reduction ansatz is

$$ds_5^2 = R^{-1} ds_4^2 + R^2 dy^2. \tag{30.22}$$

However, using the general formulas for the action for the scalar R, we can define the relation to the canonical scalar ϕ (identified with the same scalar ϕ used until now in the ekpyrotic and kinetic phases), as

$$R \equiv e^{\sqrt{\frac{2}{3}}\phi_{\text{can}}}. \tag{30.23}$$

Then substituting into the 5 dimensional metric, we obtain the ansatz (and dropping the index "can" on the canonical scalar)

$$ds_5^2 = e^{-\sqrt{\frac{2}{3}}\phi} ds_4^2 + e^{2\sqrt{\frac{2}{3}}\phi} dy^2 \,, \tag{30.24}$$

for the canonical 4 dimensional effective action (valid during the kinetic phase, when the scalar potential can be ignored)

$$S_{\rm red} = \int d^4x \sqrt{-g} \left[\frac{1}{2\kappa^2} \mathcal{R} - \frac{1}{2} (\partial_\mu \phi)^2 \right]. \tag{30.25}$$

Since ϕ is the ekpyrotic scalar, we can substitute in the 5 dimensional ansatz (30.24) the solution (30.18), (30.19), with $a_0 = 1$, $\phi_0 = 0$, and then we find

$$ds_5^2 = (-t)^{-2/3} [-dt^2 + (-t)^{2/3} d\vec{x}_3^2] + (-t)^{4/3} dy^2. \tag{30.26}$$

Making the change of variables

$$T \equiv \frac{3}{2}(-t)^{2/3}; \quad y = \frac{2}{3}\bar{y} \,, \tag{30.27}$$

we find the metric

$$ds_5^2 = -dT^2 + T^2 d\bar{y}^2 + d\vec{x}_3^2. \tag{30.28}$$

But note that by a Wick rotation, the $(T, \bar{y})$ space is a cone, with radial coordinate T and angular one $\bar{y}$, which means that the space is actually flat, except at the collision point $T = 0$ (so $t = 0$). Indeed, even without the Wick rotation, by making a coordinate transformation

$$u = T \cosh y; \quad v = T \sinh y \,, \tag{30.29}$$

we find the flat space (except at $T = 0$, which makes the coordinate transformation singular)

$$ds_5^2 = -du^2 + dv^2 + d\vec{x}_3^2. \tag{30.30}$$

If we didn't have $\vec{x}_3$, nor the compactification and orbifolding of the y coordinate, the resulting space would be called the "Milne Universe". As it is, it is called the "compactified Milne mod $\mathbb{Z}_2$", and as we saw, it describes 2 orbifold planes (where we have the branes) approaching, colliding, and then (during the Big Bang phase) receding from each other.

30.3 String Theory Model

Until now we have presented the general 2-brane model that is preferred, as we will see. But now we also present the original string theory motivation for the model, though as we will shortly see, it is by now excluded for theoretical reasons.

As we mentioned, the model is defined in heterotic M-theory, which has two "end of the world" M9-branes, situated on orbifold fixed points, at a relative distance (radion) $\pi R = \pi l_P g_s^{3/2}$. The M9-branes have each a E_8 gauge group on them, for a total $E_8 \times E_8$. We are interested in the large g_s limit (large R), and in this case we can describe the resulting theory from the point of view of the low energy supergravity, compactified to 4 dimensions. The bulk action is then the 11 dimensional supergravity action

$$S_{11} = \frac{1}{2\kappa_{11}^2} \int_{M^{11}} \sqrt{-\hat{g}} \left[\hat{\mathcal{R}} - \hat{F}_{MNPQ} \hat{F}^{MNPQ} - \frac{\sqrt{2}}{1728} \epsilon^{M_1 \dots M_{11}} \hat{A}_{M_1 M_2 M_3} \hat{F}_{M_4 \dots M_7} \hat{F}_{M_8 \dots M_{11}} \right]. \tag{30.31}$$

But there is also an anomaly action, arising from a one loop quantum worldsheet string correction (so it is of order α'), that lives only on the boundary, i.e., on the 2 M9-branes (defined by $i = 1, 2$), so that

$$S_{\text{bdy}} = \frac{1}{8\pi\kappa_{11}^2} \left(\frac{\kappa_{11}}{4\pi} \right)^{2/3} \sum_{i=1,2} \int_{\mathcal{M}_{10}^{(i)}} \sqrt{-g} \left[\text{Tr}(F^{(i)})^2 - \frac{1}{2} \text{Tr} R^2 \right]. \tag{30.32}$$

Moreover, the 6 extra coordinates (except the "eleventh one" and our 4 dimensions) must be compactified on a CY_3 space, as usual. From the general formula for nonlinear KK compactification (30.21), we obtain

$$ds_{11}^2 = \Delta^{-1/3} ds_5^2 + g_{mn} dy^m dy^n \ , \tag{30.33}$$

with $\Delta = \det g_{mn} = V^2$, where V is the volume of the CY_3, so

$$ds_{11}^2 = V^{-2/3} ds_5^2 + g_{mn} dy^m dy^n . \tag{30.34}$$

To cancel the anomaly coming from the two boundary terms, not locally but in the 4 dimensional effective field theory, we can choose the standard embedding (under KK compactification of string theory), of the gauge field on one of the branes (in one E_8 group) into the spin connection, while putting to zero the other. This means that, under dimensional reduction, the same lower dimensional field appears both in the gauge field and in part of the spin connection, and it results in having

$$\text{Tr}[F^{(1)} \wedge F^{(1)}] = \text{Tr}[R \wedge R], \qquad F^{(2)} = 0. \tag{30.35}$$

Lastly, we must introduce the M5-brane moving in the bulk, which interacts with the boundary branes via an exchange of M2-branes. One can calculate in M-theory the resulting 4 dimensional superpotential (in effective field theory),

$$W \sim e^{-cY} \ , \tag{30.36}$$

where Y is the chiral superfield corresponding to the distance between the M5-brane and the boundary M9-brane. But using the general formula for the $\mathcal{N} = 1$ supergravity coupled to chiral superfields,

$$V = e^{K/M_{\rm Pl}^2} \left[g^{i\bar{j}} D_i W \overline{D_j W} - \frac{3}{M_{\rm Pl}^2} |W|^2 \right] \ , \tag{30.37}$$

we find

$$V = -V_0 e^{-\tilde{c}\phi} \ , \tag{30.38}$$

as advertised.

Under dimensional reduction on CY_3, we obtain the effective 5 dimensional action

$$S_5 = \frac{M_5^3}{2} \int_{\mathcal{M}_5} d^5x \sqrt{-g} \left[\mathcal{R} - \frac{1}{2}(\partial\phi)^2 - \frac{3}{2} e^{2\phi} \frac{\mathcal{F}^2}{5!} \right] - 3 \sum_{i=1,2,3} \alpha_i M_5^3 \int_{\mathcal{M}_4} d^4\xi_i \left(\sqrt{-h_{(i)}} e^{-\phi} \right.$$
$$\left. - \frac{1}{4!} \epsilon^{\mu\nu\rho\sigma} \mathcal{A}_{\alpha\beta\gamma\delta} \partial_\mu X^\alpha \partial_\nu X^\beta \partial_\rho X^\gamma \partial_\sigma X^\delta \right) \ , \tag{30.39}$$

where the 3-brane terms come from the dimensional reduction of the two M9-branes on CY_3, and of the M5-brane on a 2-cycle, both leading to 3-branes in 5 dimensions.

We can then show that, for $\alpha_1 = -\alpha$, $\alpha_2 = \alpha - \beta$, $\alpha_3 = \beta$, we have a BPS solution defined by a one dimensional harmonic function (in the coordinate transverse to the branes) $D(y)$,

$$\begin{aligned} ds^2 &= D(y)(-N^2 d\tau^2 + A^2 d\vec{x}_3^2) + B^2 D^2(y) dy^2 \\ e^\phi &= B D^3(y) \\ \mathcal{F}_{0123y} &= -\alpha A^3 N B^{-1} D^{-2}(y), \quad \text{for} \quad y < Y \\ &= -(\alpha - \beta) A^3 N B^{-1} D^{-2}(y), \quad \text{for} \quad y > Y \\ D(y) &= \alpha y + C, \quad \text{for} \quad y < Y \\ &= (\alpha - \beta) y + C + \beta Y, \quad \text{for} \quad y > Y \end{aligned} \tag{30.40}$$

Here $y = Y$ is the position of the bulk brane, and the boundary branes are situated at $y = 0$ and $y = R$ (so $0 \leq y \leq R$). One can use a moduli space approximation, which means to promote the constant parameters like the position Y to fields = functions of time τ, in particular $Y(t)$, resolve the equations of motion to linear order in the perturbation, and find the effective action on the moduli space, i.e., on the space of $Y(t)$ and other fields. One will find that the effective description that we used during the kinetic phase, i.e., with a canonical scalar, is a good one.

30.4 Solving an Issue Related to the Null Energy Condition

After the initial ekpyrotic paper, it was observed in a paper by the same authors plus Seiberg that the fact that the null energy condition is satisfied presents a problem with the original ekpyrotic scenario described above, that can be solved if we consider only the end of the world branes to be the colliding ones.

The null energy condition is the most general of the energy conditions, believed to always hold. The energy conditions are conditions that replace the condition of positivity of the energy density in general relativity. Generally, the null energy condition is described as $T_{\mu\nu}n^\mu n^\nu \geq 0$, for all null n^μ vectors, i.e., $n^\mu n_\mu = 0$, but in the case of a perfect fluid, it reduces simply to

$$\rho + P \geq 0. \tag{30.41}$$

For gravity coupled to scalars (the various moduli of a compactification) and with a potential V,

$$S = \int d^4x\sqrt{-g}\left[\frac{1}{2\kappa_4^2}\mathcal{R} - \frac{1}{2}g^{\mu\nu}G_{IJ}(\phi^K)\partial_\mu\phi^I\partial_\nu\phi^J - V(\phi^K)\right], \tag{30.42}$$

which has the energy-momentum tensor

$$T_{\mu\nu} = G_{IJ}\partial_\mu\phi^I\partial_\nu\phi^J - g_{\mu\nu}\left[\frac{1}{2}g^{\alpha\beta}G_{IJ}\partial_\alpha\phi^I\partial_\beta\phi^J + V\right], \tag{30.43}$$

the null energy condition becomes

$$\rho + P \equiv T_{00} - \frac{1}{3}g^{ij}T_{ij} = G_{IJ}\dot{\phi}^I\dot{\phi}^J \geq 0. \tag{30.44}$$

The null energy condition is indeed satisfied, just because of the positivity of the kinetic term for the scalars, i.e., because of the absence of ghosts in a normal theory.

But by substituting this result in the second Friedmann equation, we obtain

$$\dot{H} = -4\pi G_N(\rho + P) = -4\pi G_N G_{IJ}\dot{\phi}^I\dot{\phi}^J \leq 0. \tag{30.45}$$

However, that means that we cannot have a reversal from a contracting phase, during which $H < 0$, to an expanding phase, during which $H > 0$, via matter satisfying the null energy condition, since it gives $\dot{H} < 0$.

In order to violate this simple theorem, we must either:

- have terms in the action that are higher order in derivatives and violate the null energy condition, or
- pass through a *singularity*, where we can jump discontinuously from $H < 0$ to $H > 0$, so where $\dot{H}$ is not well defined.

Clearly we are not in the first case, so we must be in the second. But the problem with it is that we then cannot calculate things. However, it turns out that there are two descriptions of the physics, one in which there is a singularity, which is the usual cosmology that we described, but there is another one where there are no singularities, so we can calculate things.

Consider the kinetic dominated phase, with $V \simeq 0$, near the singularity, and the FLRW metric in conformal time η, so

$$ds^2 = a^2(\eta)[-d\eta^2 + d\vec{x}_3^2]. \tag{30.46}$$

The solution for $(a(t), \phi(t))$ is rewritten as

$$\begin{aligned} a &= a(1)\eta^{1/2}; \\ \phi &= \phi(0) + \sqrt{3}\epsilon \log|\eta| \ , \end{aligned} \tag{30.47}$$

where $\epsilon = \pm 1$.

But we can now remember that the scalar ϕ parametrizes the radion, which is related to the dilaton of string theory (being the "eleventh dimension") as

$$R = R_0 e^{\sqrt{\frac{2}{3}}\phi_{\text{can}}} = l_P g_s^{2/3} = l_P e^{\frac{2}{3}\phi_{\text{dilaton}}}. \tag{30.48}$$

But that means that we can define a *string frame*, as the natural metric that appears in string theory (from string theory calculations), related to the Einstein metric by

$$g_{\mu\nu,s} = e^{\phi_{\text{dilaton}}/c} g_{\mu\nu,E} \ , \tag{30.49}$$

where

$$c = \sqrt{\frac{d-2}{2}}. \tag{30.50}$$

Indeed, for $d = 10$, we have $c = 2$, as we said in part II, whereas for $d = 4$ we have $c = 1$. The string frame action in d dimensions is

$$S_{\text{string}} = \int d^d x \sqrt{-g} e^{-c\phi_{\text{dilaton}}} [\mathcal{R}[g_s] + c^2 g^{\mu\nu} \partial_\mu \phi_{\text{dilaton}} \partial_\nu \phi_{\text{dilaton}}]. \tag{30.51}$$

Defining the string time τ_s by

$$d\tau_s = a_s d\eta \ , \tag{30.52}$$

we find the string frame solution (we leave it as an exercise to check it is satisfied)

$$\begin{aligned} a_s &= a_s(1)|\tau_s|^{\frac{\epsilon}{\sqrt{3}}} \\ \phi_{\text{dilaton}} &= \phi_{\text{dilaton}}(0) + (\epsilon\sqrt{3} - 1)\log|\tau_s| \\ \epsilon &= \pm 1. \end{aligned} \tag{30.53}$$

Here $\epsilon = \pm 1$ is used to compare the ekpyrotic case with a previously considered case.

$\epsilon = -1$ corresponds to the "pre-Big Bang" scenario previously considered by Veneziano in string theory. However, as we can easily see, it corresponds to strong coupling near the Big Crunch/Big Bang,

$$g_s = e^{\phi_{\text{dilaton}}} \to \infty \quad \text{for} \quad t \to 0- . \tag{30.54}$$

That means that in order to describe what happens at the cosmological singularity, we must make assumptions about strongly coupled string theory, which are of course not reliable.

$\epsilon = +1$ on the other hand corresponds to the ekpyrotic case considered in this chapter, and as we can easily see has instead weak coupling at the cosmological singularity,

$$g_s = e^{\phi_{\text{dilaton}}} \to 0 \quad \text{for} \quad t \to 0- \tag{30.55}$$

That means that on the contrary, now we can trust the physics at the cosmological singularity, since it is classical (zero coupling).

But if there is a singularity, how can we trust anything that goes through it not to be modified? This is needed, since we need to consider perturbations, which are generated during the ekpyrotic phase.

The answer lies in the fact that there is another frame for gravity, besides the Einstein and string frames, in which physics is non-singular. Indeed, consider the metric frame

$$\bar{g}_{\mu\nu} \equiv e^{-\frac{\phi_{\text{dilaton}}}{\sqrt{3}}} g_{\mu\nu} , \tag{30.56}$$

which leads to the solution

$$\begin{aligned} \bar{a}(t) &= A = \text{constant} \\ e^{\frac{\phi_{\text{dilaton}}}{\sqrt{12}}} &= B|t|^{1/2}. \end{aligned} \tag{30.57}$$

Therefore as we can see, while again we have the string coupling $g_s = e^{\phi_{\text{dilaton}}}$ vanishing at the cosmological singularity, there is in fact no singularity in the metric, since the spatial metric is constant.

The existence of these coordinates should have been guessed from the brane picture. Indeed, the original description of the cosmology was in terms of a collision of branes, and clearly in that picture there is nothing singular happening with the 3 dimensional spatial volume at the collision point: it still looks finite; in fact, the whole 3 dimensional space on the branes doesn't look like it changes at all near the transition, just like in the $\bar{a}(t)$ coordinates above.

The important observation is that the mechanism described here *only* works if the ekpyrotic scalar ϕ is the radion, which is identified in heterotic M-theory with the dilaton ϕ_{dilaton}. That means that we need to only have the boundary branes colliding. If we would have instead the bulk brane colliding with the boundary brane, at fixed

interval (radion) distance R, so fixed dilaton, the mechanism described here would not work. That eliminates the calculable scenario which was the original motivation for ekpyrosis.

30.5 Fluctuations in Ekpyrotic Models

The most important issue nowadays, given that we can measure very precisely the CMBR spectrum, is to find the spectrum of fluctuations.

- One possible assumption is that the full metric, including fluctuations, goes unchanged through the bounce from Big Crunch to Big Bang. There have been many papers claiming this, with various variants for it.
- But there have been also papers claiming that there are dynamical issues happening at the cosmological singularity. I believe that the issue is not yet settled.

If the first assumption is made, then, defining the quantity

$$\epsilon \equiv \frac{3}{2}(1+w) \gg 1 \,, \tag{30.58}$$

after a long calculation that will not be reproduced here, one finds the tilt of the power spectrum coming in the Newtonian potential Φ as

$$n_\Phi - 1 = 1 - \left|\frac{\epsilon+1}{\epsilon-1}\right| \ll 1. \tag{30.59}$$

But we need to calculate the spectrum of the curvature perturbations ζ, as we saw in part I. It turns out that this needs a mixing of modes that is model dependent, so the result above is modified in general. Note that we get $n-1 \ll 1$, like for inflation (though we still need to check that the same happens for ζ fluctuations).

But the thing that most distinguishes the ekpyrotic scenario from the inflationary one is the spectrum of tensor fluctuations. One finds that this spectrum has the tilt

$$n_T = 3 - \left|\frac{\epsilon-3}{\epsilon-1}\right| \simeq 2 \,, \tag{30.60}$$

with contrast to inflation, where we find also $n_T \simeq 1$ (a scale invariant spectrum). Thus we find a very blue spectrum. But since we haven't even seen tensor perturbations, much less measure the tilt of their spectrum, this distinguishing feature is unlikely to be tested in the very near future.

Important Concepts to Remember

- The ekpyrotic scenario is based on two end-of-the-world branes, moving towards each other and colliding.
- Unlike inflation, everything relevant happens before the Big Bang, in a contracting phase with a negative potential.
- The original ekpyrotic model was in heterotic M-theory, with a bulk M5-brane colliding with one of the M9-branes carrying E_8 gauge groups.
- Fluctuations are on the branes, before the Big Bang (collision), and the horizon problem is fixed because of the cohesion of the branes before the collision, and their pre-Big Bang causal connection.
- There is an effective field theory description in terms of $\phi(\vec{x}, t)$, the interbrane separation.
- For the contracting ekpyrotic phase, we need $V = -V_0 e^{-c\phi}$, with $c \gg 1$, which leads to $a(t) \propto (-t)^p$, $p \ll 1$, corresponding to an effective equation of state with $w \gg 1$.
- After the ekpyrotic phase, we have a kinetic phase, dominated by a kinetic term, with $w = 1$ (stiff matter), and after it we reach the Planck scale.
- In this late stage the 5 dimensional picture prevails, and we obtain a "compactified Milne Universe mod $\mathbb{Z}_2$", for 2 orbifold planes colliding and receding.
- The null energy condition (NEC) $\rho + P \geq 0$, satisfied by a scalar field with a potential, implies via the Friedmann equation that $\dot{H} = -4\pi G_N(\rho + P) \leq 0$, i.e., we cannot go from a contracting ($H < 0$) Universe to an expanding ($H > 0$) Universe, unless we pass through a singularity or violate the NEC.
- While we pass through a singularity, there is a gravitational frame in which things are smooth (constant scale factor as a function of time). With a choice $\epsilon = +1$, we even have $g_s \to 0$ at the singularity, so we can pass through it.
- Like for inflation, scalar fluctuations have $n_\Phi - 1 < 0$ and $\ll 1$, however tensor fluctuations have $n_T - 1 \simeq 1$, unlike inflation, where it is also $\ll 1$.

Further reading: The original paper on the ekpyrotic scenario is [52]. The analysis of the transition to the Big Bang and the necessity of 2-brane model was done in [53]. A good review of the ekpyrotic (and cyclic) scenarios is [54].

Exercises

(1) Check that

$$\begin{aligned} a(t) &= (-t)^p \\ \phi_i &= \frac{2}{c_i} \log\left(-t\sqrt{\frac{c_i^2 V_i}{2}}\right) \\ p &= \sum_i \frac{2}{c_i^2} \end{aligned} \tag{30.61}$$

solves the equations of motion of the cosmology with potential $V = \sum_i V_i(\phi_i)$, where

$$V_i(\phi_i) = -V_i e^{-c_i \phi_i}. \tag{30.62}$$

(2) Check that the solution

$$\begin{aligned} ds^2 &= D(y)(-N^2 d\tau^2 + A^2 d\vec{x}_3^2) + B^2 D^2(y) dy^2 \\ e^{\phi} &= B D^3(y) \\ \mathcal{F}_{0123y} &= -\alpha A^3 N B^{-1} D^{-2}(y), \quad \text{for} \quad y < Y \\ &= -(\alpha - \beta) A^3 N B^{-1} D^{-2}(y), \quad \text{for} \quad y > Y \\ D(y) &= \alpha y + C, \quad \text{for} \quad y < Y \\ &= (\alpha - \beta) y + C + \beta Y, \quad \text{for} \quad y > Y \end{aligned} \tag{30.63}$$

satisfies the ϕ and $\mathcal{F}$ equations of motion for the action

$$\begin{aligned} S_5 = \frac{M_5^3}{2} \int_{\mathcal{M}_5} d^5x \sqrt{-g} \left[\mathcal{R} - \frac{1}{2}(\partial\phi)^2 - \frac{3}{2} e^{2\phi} \frac{\mathcal{F}^2}{5!} \right] - 3 \sum_{i=1,2,3} \alpha_i M_5^3 \int_{\mathcal{M}_4} d^4\xi_i \left(\sqrt{-h_{(i)}} e^{-\phi} \right. \\ \left. - \frac{1}{4!} \epsilon^{\mu\nu\rho\sigma} \mathcal{A}_{\alpha\beta\gamma\delta} \partial_\mu X^\alpha \partial_\nu X^\beta \partial_\rho X^\gamma \partial_\sigma X^\delta \right), \end{aligned} \tag{30.64}$$

for $\alpha_1 = -\alpha$, $\alpha_2 = \alpha - \beta$, $\alpha_3 = \beta$.
(3) Check that the string frame solution

$$\begin{aligned} a_s &= a_s(1) |\tau_s|^{\frac{\epsilon}{\sqrt{3}}} \\ \phi &= \phi(0) + (\epsilon\sqrt{3} - 1) \log |\tau_s| \\ \epsilon &= \pm 1 \end{aligned} \tag{30.65}$$

satisfies the equations of motion of the action

$$S = \int d^d x \sqrt{-g_s} e^{-c\phi} (R[g_s] + c^2 g^{\mu\nu} \partial_\mu \phi \partial_\nu \phi), \tag{30.66}$$

for

$$c = \sqrt{\frac{d-2}{2}}. \tag{30.67}$$

(4) Show that the 5 dimensional action (30.39) is obtained by dimensional reduction on CY_3 from the 11 dimensional supergravity action (30.31).

Chapter 31
The Cyclic and New Ekpyrotic Scenarios

In this chapter we will study variants of the original ekpyrotic scenario, which are introduced in order to solve some problems with the original model. These are the cyclic model of Steinhardt and Turok, and the new ekpyrotic model of Buchbinder, Khoury and Ovrut.

31.1 The Cyclic Model

The cyclic model is described within 4 dimensional effective field theory, by a potential $V(\phi)$ that has the following properties:

- it must go asymptotically, for $\phi \to +\infty$, to a constant V_0, giving the "cosmological constant", or rather dark energy, today. Only for small ϕ close to zero, the potential bends down and goes through zero. Thus for $\phi > 0$ we are in the standard cosmology, dominated asymptotically by dark energy, with equation of state $w \simeq -1$.
- for negative ϕ, until a certain value, we must have an ekpyrotic phase, with $w \gg 1$, so the potential must have a steep (exponential) drop, until a minimum value of $-V_{\text{end}}$.

Combining these two features, we have the following potential:

$$V = V_0(1 - e^{-c\phi}) \,, \tag{31.1}$$

where from the ekpyrotic condition of $w \gg 1$, as we saw in the previous chapter, we need $c \gg 1$. In fact, from making contact with experiments we will see that we need $c \gtrsim 10$.

- Finally, as we said last time, the potential cannot increase indefinitely, and near the Planck scale we should have kinetic energy domination. That means that the potential should reverse, forming a well shape with minimum at $-V_{\text{end}}$, and go

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_31

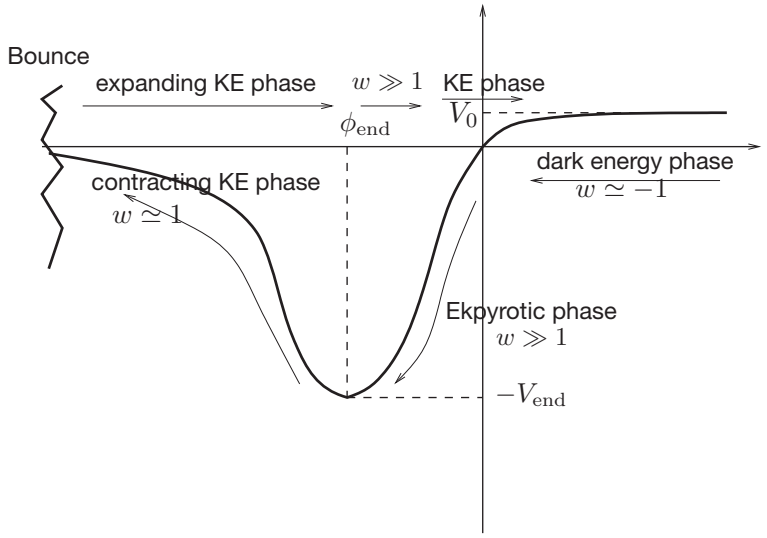

Fig. 31.1 The potential for the cyclic model, with a sketch for the various phases of evolution in it

back to zero. The kinetic domination phase applies for this upward branch of the potential, with equation of state $w \simeq 1$ (stiff matter). At the end, when one reaches zero, the bounce from Big Crunch to Big Bang should happen.

The potential is sketched in Fig. 31.1.

So the picture of the cyclic cosmology is as follows. Cosmology happens in cycles, expansion to a maximum, followed by contraction to zero (Big Crunch), followed by Big Bang and expansion. We start in a dark energy phase, with kinetic energy that is just shy of the one needed to reach the asymptotic value V_0, and after a maximum, we start to contract and go back to smaller ϕ. For $\phi < 0$ we enter a contracting ekpyrotic phase, followed by the contracting kinetic phase, followed by a bounce. Then we start expanding in a kinetic phase, followed by a quick ekpyrotic phase (that can be neglected), and then we are back to standard cosmology: radiation domination, followed by dark energy domination. And the cycle repeats.

We will see that the entropy increases in each cycle, but so does the volume, which increases exponentially over a cycle, so in effect, we are left with the same entropy density.

The scalar ϕ corresponds in the original string theory construction to the distance between the colliding brane, and the bounce to the collision between branes (that generates entropy).

In the brane picture, we have 2 branes, for a bulk dimensionally reduced metric

$$ds_{11}^2 = V^{-2/3} ds_5^2 + V^{1/3} \tilde{g}_{mn} dy^m dy^n. \tag{31.2}$$

Here note that $V = \sqrt{\det g} = \Delta^{1/2}$ is the volume of the internal 6 dimensional space, hence the $\Delta^{-1/(d-2)} = \Delta^{-1/3} = V^{-2/3}$ prefactor in front of the 5 dimensional metric. We have also taken out of the metric a factor of $V^{1/3} = \det g^{1/6}$, so that the remaining metric has unit determinant, $\det \tilde{g}_{mn} = 1$. Moreover, $V = e^{\phi}$ is the 4 dimensional dilaton. Then the effective action in 5 dimensions is

$$S_5 = \frac{1}{2\kappa_5^2}\int d^5x\sqrt{-g}\left[\mathcal{R} - \frac{\partial_\mu V \partial^\mu V}{V^2} - \frac{6\alpha^2}{V^2}\right] + \frac{12\alpha}{2\kappa_5^2}\left[-\int_{y=+1} d^4x\sqrt{-g}\frac{1}{V} + \int_{y=-1} d^4x\sqrt{-g}\frac{1}{V}\right]. \tag{31.3}$$

Note that the potential term is now, once we canonically normalize V to ϕ, $\sim e^{-\sqrt{2}\phi}$, and that the brane sources are at $y = \pm 1$. We obtain a domain wall solution,

$$\begin{aligned} ds_5^2 &= h^{2/5}(y)[A^2(-d\tau^2 + d\vec{x}^2) + B^2 dy^2] \\ V &= Bh^{6/5}(y) \\ h(y) &= 5\alpha y + C\ , \end{aligned} \tag{31.4}$$

and symmetric with respect to $y = \pm 1$, so that it actually satisfies the Israel junction conditions at these branes.

In this way we have obtained two static branes, but we can use the usual moduli space approximation for the brane solution, where we make the parameters in the (solitonic) solution, including the position of the (solitonic) brane, as functions of time, $A, B, C \to A(\tau), B(\tau), C(\tau)$. Then as usual, we don't have anymore a solution, but we can re-solve the equations of motion to linear order in the velocities (in the time dependence), and find the *motion on moduli space*, and its associated effective action on moduli space. In this way, we find the colliding brane solution and its effective field theory.

Above we have obtained the exponential term in the potential (like we also had in the previous chapter). But the positive constant part of the potential V_0, which would mean a de Sitter background, is harder to obtain.

During the ekpyrotic phase, we have a very steep drop in potential, and we generate perturbations, but unlike the case of inflation, we do so in a "fast-roll" approximation, the opposite to the "slow-roll" of inflation. For a single field ekpyrosis, the spectral tilt of generated fluctuations can be found to be

$$n_s - 1 \simeq -4(\epsilon + \eta)\ , \tag{31.5}$$

where however now ϵ and η are "fast-roll" parameters

$$\epsilon = \frac{1}{M_{\rm Pl}^2}\left(\frac{V}{V'}\right)^2$$

$$\eta = 1 - \frac{V''V}{V'^2}. \tag{31.6}$$

The "fast-roll" approximation means that $\epsilon, \eta \ll 1$, so indeed we obtain $n_s - 1 \ll 1$, as it observed experimentally. Moreover, using the potential above, we find

$$n_s - 1 = -\frac{4}{c^2}\ , \tag{31.7}$$

and from the experimental value $n_s - 1 \simeq -0.04$, we find indeed

$$c \gtrsim 10. \tag{31.8}$$

Note that for *two-field ekpyrosis*, one obtains instead

$$n_s - 1 \simeq +4\epsilon - 4\eta. \tag{31.9}$$

We now turn to the evolution of the space, i.e., the evolution of $a(t)$ and $H(t)$.

- During the *dark energy phase*, $a(t)$ will grow a certain number of e-folds, by definition $N_{\rm DE}$, so $a(t)$ increases by $e^{N_{\rm DE}}$, whereas $H(t)$ remains approximately constant.
- During the *ekpyrotic phase*, since, as we saw last time,

$$V(\phi(t)) \propto \frac{1}{t^2}\ , \tag{31.10}$$

and the time evolution is

$$a(t) \propto (-t)^p, \quad H(t) \propto \frac{1}{t}\ , \tag{31.11}$$

we obtain

$$\begin{aligned} a(t) &\propto V(\phi)^{-\frac{p}{2}} \\ H(t) &\propto V(\phi)^{1/2}\ , \end{aligned} \tag{31.12}$$

so $a(t)$ shrinks by

$$\left|\frac{V_{\rm end}}{V_{\rm beg}}\right|^{-p/2} \sim \mathcal{O}(1)\ , \tag{31.13}$$

where this factor is not large, since $p \ll 1$. On the other hand, $H(t)$ grows by

$$\left|\frac{V_{\rm end}}{V_{\rm beg}}\right|^{1/2} \equiv e^{N_{ek}}\ , \tag{31.14}$$

where in the last equality we have defined the number of e-folds of ekpyrosis.

- After the ekpyrotic phase, we have a period of kinetic domination, starting at $V_{\rm end}$ (the bottom of the potential well), followed by the bounce, that creates energy from collisions (*reheats* the Universe), followed by kinetic phase again, and a short (and irrelevant, since it will not generate more perturbations during expansion) ekpyrotic phase, until *thermalizing* at a reheat temperature T_r (so that $\rho_{\rm reheat} \sim T_r^4$), once we reach positive potential. This then transitions into the standard radiation dominated cosmology. In effect then, we start with $\rho \sim |V_{\rm end}| \sim T^4$, and we end with $\rho \sim T_r^4$. Since during this (mostly) kinetic dominated phase, we have $w \gg 1$, so $\rho \propto a^{-6} \propto t^{-2}$ (in general $\rho \propto a^{-3(1+w)}$), and therefore $a(t)$ increases by

$$\left(\frac{|V_{\rm end}|^{1/4}}{T_r}\right)^{2/3} \equiv e^{\frac{2}{3}\gamma_{\rm KE}}. \tag{31.15}$$

  On the other hand, $H(t) \propto 1/t \propto \rho^{1/2} \propto T^2$, so shrinks by

$$\left(\frac{|V_{\rm end}|^{1/4}}{T_r}\right)^{-2} = e^{-2\gamma_{\rm KE}}. \tag{31.16}$$

- During the standard radiation dominated cosmology, $\rho \propto a^{-4} \propto 1/t^2 \propto T^4$, so $a(t)$ increases by

$$\frac{T_r}{T_0} \equiv e^{N_{\rm rad}}\ , \tag{31.17}$$

  where T_0 is the CMBR temperature today. Also, $H(t) \propto 1/t \propto \sqrt{\rho} \propto T^2$, so shrinks by

$$e^{-2N_{\rm rad}}. \tag{31.18}$$

  Putting everything together, we find that $a(t)$ grows in a cycle by

$$\exp\left(N_{\rm DE} + \frac{2}{3}\gamma_{\rm KE} + N_{\rm rad}\right). \tag{31.19}$$

  On the other hand, $H(t)$ is modified by

$$\frac{\sqrt{|V_{\rm end}|}}{\sqrt{V_{\rm beg}}}\frac{T_r^2}{\sqrt{|V_{\rm end}|}}\frac{T_0^2}{T_r^2} = \frac{T_0^2}{V_{\rm beg}^{1/2}}. \tag{31.20}$$

But this ratio is approximately one, as indeed it should be, since over a cycle, $H(t)$ must return to the same value, the one given by the dark energy V_0. In another way, this gives a constraint on N_{ek}, since the total number of e-folds in $H(t)$ is $N_{ek} - 2(\gamma_{\rm KE} + N_{\rm rad})$:

$$N_{ek} \simeq 2(\gamma_{\rm KE} + N_{\rm rad}). \tag{31.21}$$

We have some constraints on the reheat temperature T_r. First, like in the case of inflation, T_r must be larger than the Big Bang Nucleosynthesis temperature of about 1MeV, which is the first time that we know precisely what happens in cosmology (we know nuclear reactions, and we can find that changing things would mess up the primordial abundances). Second, we know it must be smaller than the GUT scale of about 10^{16} GeV, to avoid the monopole problem (the appearance of undiluted solitons coming from the GUT phase transition). All in all,

$$1\text{MeV} < T_r < 10^{16}\,\text{GeV}. \tag{31.22}$$

But another constraint appears purely from the condition to have radiation domination start only after the crossing of the potential well. Since during the kinetic dominated phase (assuming that roughly the same exponential potential is valid), as we saw last time,

$$\phi(t) \propto \sqrt{\frac{2}{3}}\ln(t) \Rightarrow V(\phi(t)) \propto t^{-c\sqrt{\frac{2}{3}}}\,, \tag{31.23}$$

and assuming that the potential at the beginning of the phase is of the order of $|V_{\text{end}}|$ (we work in order of magnitude only anyway) and after the crossing is of about V_0, we obtain

$$t_{\text{cross}} = t_{\text{end}}\left(\frac{|V_{\text{end}}|}{V_0}\right)^{\sqrt{\frac{3}{2}}\frac{1}{c}}. \tag{31.24}$$

Then, from the condition that reheating happens (just) after crossing, $t_r \gtrsim t_{\text{cross}}$, and from the fact that $t_r \sim H_r^{-1}$ ($H \propto 1/t$), and $t_r \sim 1/\sqrt{\rho} \sim T_r^{-2}$, we find

$$T_r \lesssim |V_{\text{end}}|^{1/4}\left|\frac{V_0}{V_{\text{end}}}\right|^{\sqrt{\frac{3p}{16}}}. \tag{31.25}$$

As an example, considering a reasonable $|V_{\text{end}}| \sim (10^{15}\,\text{GeV})^4$, and a $p \sim 10^{-3}$, we find

$$T_r \lesssim 10^{12}\,\text{GeV}. \tag{31.26}$$

31.2 The New Ekpyrotic Cosmology

New ekpyrotic cosmology does not attempt to modify the basic scenario of the ekpyrotic case (unlike the cyclic cosmology), but rather attempts to deal with the fact that during the kinetic dominated phase we must switch from a contracting ($H < 0$) Universe to an expanding one ($H > 0$), and as we saw, there is a simple theorem saying we can either have a singularity, or violate the null energy condition.

In the ekpyrotic and cyclic cosmologies, the assumption is that it can happen since we have a singularity, but on the other hand we can calculate what happens to the spectrum of fluctuations since it can go untouched through it, due to the existence of a frame where there is no singularity, and where the string coupling goes to zero at the transition. But various points in that reasoning have been contested, hence the need for an alternative, and the logical alternative is to consider a violation of the null energy condition.

Now a random violation of the null energy condition is bad, since the NEC is assumed to hold in some generality. But one example of a better violation is in the system known as *ghost condensate* (since there is the potential to have a ghost, which of course would also be bad; but that is avoided).

The action one takes is a general function of the kinetic variable

$$X \equiv -\frac{1}{2m^4}(\partial\phi)^2 > 0 \ , \tag{31.27}$$

so

$$S = \int d^4x\sqrt{-g}P(X) \ , \tag{31.28}$$

where $P(X)$ is an arbitrary function. Note that a case where $P(X) \simeq A + BX$ at small X, with various higher derivative corrections, has been considered before, going under the name of k-essence if it tries to define dark energy, or k-inflation (or k-flation) for the application to inflation. This would be the case of the DBI action, which is actually

$$P(X) = \sqrt{1+X} \simeq 1 + X + \cdots \tag{31.29}$$

But here we will not be interested in that case. For a cosmology with uniform field ϕ and scale factor, so $a = a(t)$, $\phi = \phi(t)$, the equation of motion coming from the above action is

$$\frac{d}{dt}(a^3 P_{,X}\dot{\phi}) = 0. \tag{31.30}$$

But we want to consider the case that $P(X)$ has *minimum* at some finite X, which can be taken to be (by a redefinition of fields and of m) $X_0 = 1/2$. Then $X_0 = 1/2$ is an *exact* solution of the above equation of motion (since $P_{,X}(X_0) = 0$), and it corresponds to

$$\phi = -m^2 t. \tag{31.31}$$

Note that the sign is arbitrary, and we have chosen the minus for convenience.

The energy-momentum tensor coming from the above action, if $M = m$, is

$$T_{\mu\nu} = g_{\mu\nu}M^4 P(X) + P_{,X}\partial_\mu\phi\partial_\nu\phi \ , \tag{31.32}$$

and it leads to the energy density and pressure (in Minkowski space)

$$\rho = T_{00} = M^4(2P_{,X} - P)$$
$$P = T_{ii} = M^4 P(X). \tag{31.33}$$

At the extremum X_0, where $P_{,X}(X_0) = 0$, we have $P = -\rho$, so $w = -1$, just like a cosmological constant. This is not surprising, since then we can just approximate the action as $S = \int d^4x\sqrt{-g}M^4 P(X_0)$, which is indeed a cosmological constant.

We now consider fluctuations around this minimum. Define

$$\phi(t, \vec{x}) = -m^2 t + \pi(t, \vec{x}) \ , \tag{31.34}$$

where $\pi(t, \vec{x})$ is the fluctuation. Then

$$X \simeq X_0 - \frac{\dot{\pi}}{m^2} + \frac{\dot{\pi}^2 - (\vec{\nabla}\pi)^2}{2m^4} + \mathcal{O}(\pi^3) \ , \tag{31.35}$$

where we have used the fact that $X_0 = 1/2$. Substituting in the action, we find

$$S \simeq \int d^4x\sqrt{-g}\Big[M^4 P(X_0) - P_{,X}(X_0)M^4\frac{\dot{\pi}}{m^2}$$
$$+ \big(P_{,X}(X_0) + 2X_0 P_{,XX}(X_0)\big)\dot{\pi}^2 - P_{,X}(X_0)(\vec{\nabla}\pi)^2\Big] + \cdots \tag{31.36}$$

But at the minimum $P_{,X}(X_0) = 0$, we have no ghosts: the leading kinetic term $\dot{\pi}^2$ has the usual sign (the sign would be opposite, and we would have a ghost, at a maximum of $P(X)$). We must then consider a higher order action for the gradient of π (one not encoded in $P(X)$), namely

$$\mathcal{L}_{\text{gradient}} = -\frac{1}{M'^2}(\vec{\nabla}^2\pi)^2 \ , \tag{31.37}$$

besides the kinetic term

$$\mathcal{L}_{\text{kinetic}} = P_{,XX}(1/2)\dot{\pi}^2 \ , \tag{31.38}$$

leading to a dispersion relation for plane waves

$$\omega^2 \sim \frac{k^4}{M'^2}. \tag{31.39}$$

Also at the minimum, expanding

$$P_{,X} \simeq P_{,X}(X_0) - P_{,XX}(X_0)\frac{\dot{\pi}}{m^2} + \cdots \ , \tag{31.40}$$

and adding a potential V (that can absorb the constant $-P(X_0 = 1/2)$), we obtain near the minimum the energy density and pressure

$$\rho \simeq -\frac{KM^4\dot{\pi}}{m^2} + V$$
$$P \simeq -V \ , \tag{31.41}$$

where

$$K \equiv P_{,XX}(1/2). \tag{31.42}$$

Then the second Friedmann equation, which was $M_{\rm Pl}^2\dot{H} = -1/2(\rho + P)$, and which by the NEC should have been negative, is now

$$M_{\rm Pl}^2\dot{H} = \frac{1}{2}\frac{KM^4\dot{\pi}}{m^2} \ , \tag{31.43}$$

and can have either sign, depending on $\dot{\pi}$. In particular, if $\dot{\pi} > 0$, we have $\dot{H} > 0$, unlike standard cosmology.

Then the KG equation for the scalar (in the presence of a potential) becomes

$$\frac{M^4}{m^4}\frac{1}{a^3}\partial_t(a^3 P_{,X}\dot{\phi}) = -V_{,\phi} \ , \tag{31.44}$$

which becomes further

$$\ddot{\pi} + 3H\dot{\phi} = -\frac{V_{,\phi}}{K}\frac{m^4}{M^4}. \tag{31.45}$$

We consider the same field ϕ to be the ghost condensate field, and the ekpyrotic field. Then, during the ekpyrotic phase, we want to have the usual situation, so

$$M^4 P(X) \simeq m^4 X = -\frac{1}{2}(\partial_\mu\phi)^2 \ , \tag{31.46}$$

whereas during the kinetic dominated phase we need to have a ghost condensate phase, so

$$P(X) \simeq \frac{K}{2}\left(X - \frac{1}{2}\right)^2. \tag{31.47}$$

Of course, it could be that $P(X) \simeq X$ (linear) at small X, and then we have the minimum at 1/2, but that would mean that there would be a maximum in between 0 and 1/2, and a maximum would mean we do have a perturbative ghost (as we already argued), which is bad.

So instead, it must be that *after* the minimum at $X_0 = 1/2$, $P(X)$ starts growing almost linearly, which is the branch that is associated with ekpyrosis. In fact, in order to have a violation of the NEC as we want, we must have a potential that is approximately

$$V(\phi) \simeq \alpha\Lambda^4\left(1 - \beta\frac{\Lambda^2}{m^2}\frac{\phi}{M_{\rm Pl}}\right) \ , \tag{31.48}$$

with $\alpha, \beta > 0$. The constant part is as it should be from just $P(X_0)$.

This ghost condensation seems then to solve most of our problems with the ekpyrotic phase, however it is unclear how one would obtain it in string theory, or some UV consistent theory. So we have traded a solution for the cosmology to not being able to obtain the model as an effective field theory, at least not in an obvious way.

There are conditions on the model one could impose. Noting that $H(t) \propto V^{1/2}$ (from the Friedmann equation), during ekpyrosis we find

$$e^{\mathcal{N}} \equiv \frac{H_{\rm min}}{H_{\rm ekp}} = \sqrt{\frac{V_{\rm min}}{V_{\rm beg}}}. \tag{31.49}$$

One can in fact show that we need

$$e^{-2\mathcal{N}}|V_{\rm min}| = V_{\rm beg} \gg m^4\ , \tag{31.50}$$

and also $|V_{\rm min}| \ll M^4 K/p$, leading to a total interval for the potential at the minimum,

$$m^4 e^{2\mathcal{N}} \ll |V_{\rm min}| \ll \frac{M^4 K}{p}. \tag{31.51}$$

Important Concepts to Remember

- The cyclic model is based on a potential that goes asymptotically to a small constant $V_0 > 0$ as $\phi \to \infty$, and for negative ϕ has a steep exponential drop until a minimum value of $-V_{\rm end}$, and then goes back to zero at even more negative ϕ. For the first two stages, a good choice is $V = V_0(1 - e^{-c\phi})$.
- Cosmology happens in cycles, with one cycle being a dark energy phase, with expansion to a maximum, then contraction, then contracting ekpyrotic phase, then kinetic phase, then bounce. Then expanding kinetic phase, quick ekpyrotic phase, and back to standard cosmology: RD, MD, dark energy phase.
- Entropy and volume increase exponentially over a cycle, with a steady entropy density.
- During the ekpyrotic phase, we have fluctuations being generated in a "fast-roll" approximation, with parameters ϵ, η, such that $n_s - 1 = -4(\epsilon + \eta)$ for single-field ekpyrosis.
- At the end of the short expanding ekpyrotic phase, and entering into the RD era, we have a thermalization at a reheating temperature T_r, which must be > 1MeV, and has an upper bound as well.
- New ekpyrotic cosmology breaks the NEC through a ghost condensate.
- A ghost condensate has an action $S = \int \sqrt{-g} P(X)$, where $X = -(\partial\phi)^2/2m^4$, that has a minimum at $X_0 = 1/2$, i.e., $\phi = -m^2 t$. After the minimum, it starts growing almost linearly, for an usual kinetic term.

Further reading: The cyclic scenario was defined in [55]. The new ekpyrotic scenario was defined in [56]. A good review of the ekpyrotic and cyclic scenarios is [54].

Exercises

(1) Calculate the evolution of the "fast-roll" parameters ϵ, η in the full potential $V = V_0(1 - e^{-c\phi})$, and evaluate the limits for the ekpyrotic and dark energy phases. Repeat for the "slow-roll" inflation parameters ϵ, η.
(2) Consider a ghost condensate with $P(X) = (X - 1/2)^2$. Calculate the equation of state $w(X)$ at arbitrary X.
(3) Show that the null energy condition can be violated if

$$V(\phi) \simeq \alpha\Lambda^4 \left(1 - \beta \frac{\Lambda^2}{m^2} \frac{\phi}{M_{\rm Pl}}\right) , \tag{31.52}$$

with $\alpha, \beta > 0$.
(4) For the case at exercise 2, calculate the time evolution of the scale factor, $a(t)$.

Chapter 32
String Gas Cosmology: Basics and Brandenberger-Vafa Scenario

In this chapter we address the first truly stringy cosmological model, the string gas model, started by Brandenberger and Vafa. The reason for the model starts with a very simple observation: If inflation lasts for more than 70 e-folds (which admitedly is not necessary, we need more than 50–60 e-folds, but restricting $N_e < 70$ would not be needed for experimental reasons, only for calculational reasons, which seems unreasonable), then current cosmological scales, $\sim H_0^{-1}$, that leave an imprint in the CMBR, were less than Planck size $l_{\rm Pl}$ at the initial time. But then we *cannot* trust the effective field theory of inflation, and we would need a full quantum gravity theory, like string theory.

The problem is that we don't have a nonperturbative quantum gravity theory (string theory), whereas for the sub-Planckian regime, this is exactly what we would need. The string gas model, and its generalization, the brane gas model, to be described in the next chapter, is an attempt to formulate a simple string theory model that still preserves some nonperturbative information, specifically T-duality.

Indeed, T-duality says that small scales are the same as large scales in string theory. For a compactified dimension, physics with radius R is the same as physics with radius α'/R, valid even non-perturbatively. If therefore we introduce this T-duality information into the model, we have a model that has some chance of working. The T-duality invariance is in fact related to the existence of a *Hagedorn temperature*, that acts like a phase transition temperature.

32.1 Hagedorn Temperature

The existence of the Hagedorn temperature is due to the exponential increase in the degeneracy of string states (the number of string states with a given energy) at large energy, which is found to be

$$d(m) \simeq e^{4\pi\sqrt{\alpha'}m} \ , \tag{32.1}$$

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_32

where the energy is

$$E_h = m = \frac{4}{\sqrt{\alpha'}}(h-1)\ , \tag{32.2}$$

and h is the conformal weight of the state.

But the thermal partition function at temperature $T = 1/\beta$ is

$$Z(\beta) = \sum_h e^{-\beta E_h} d(m)\ , \tag{32.3}$$

which in the large m limit becomes the integral

$$Z(\beta) = \int_0^\infty dm e^{4\pi\sqrt{\alpha' m}} e^{-\beta m}\ , \tag{32.4}$$

which diverges if

$$T > T_H = \frac{1}{4\pi\sqrt{\alpha'}}. \tag{32.5}$$

Here T_H is the *Hagedorn temperature*, where the diverging partition function signals a phase transition, to a system described in terms of different degrees of freedom. Note that at $T = T_H$,

$$\beta = \beta_H = 4\pi\sqrt{\alpha'} \Rightarrow d(m) = e^{b_H m}. \tag{32.6}$$

The exponential increase in the number of states, signaling a phase transition, is also observed in QCD, close to the deconfinement phase transition (and also related to the degeneracy of states of the QCD string).

But the existence of the Hagedorn temperature is related to T-duality invariance, since we need to consider compact directions in order to find β_H. If we consider some directions uncompactified, we find that the specific heat of the system is negative, $C < 0$, whereas to find $C > 0$, we need to consider all directions compactified.

If we consider a fully compactified (9-dimensional in the case of the superstring) space, the density of states at large energy is found to be of the type

$$d(E) = c\frac{e^{\beta_H E}}{E}. \tag{32.7}$$

A (long) string calculation finds indeed this form, with $c = 1$, density of states at level n of

$$d_n = (2n)^{-\frac{11}{4}} e^{2\pi\sqrt{2n}}\ , \tag{32.8}$$

where the energy of a string state at large level n is $M^2 = E^2 \simeq n/\alpha'$, so we obtain the previous result. We also can obtain positive specific heat, $C > 0$.

The assumptions for the calculation are:

- homogenous fields, i.e., functions of time t only.
- adiabatic approximation, or small time t dependence, which means that we can ignore higher derivatives, i.e., higher α' corrections (quantum worldsheet string corrections).
- weak coupling $g_s \ll 1$, which means that we can ignore quantum string corrections as well.
- all dimensions to be compactified, as we saw.

32.2 3 Large Spatial Dimensions

Given that we need to have all dimensions compactified for $C > 0$, but we know there are 3 very large, apparently uncompactified, dimensions, the question is, how this came to be? It seems natural to start at the Planck scale, with all dimensions compactified with the same scale. Through the cosmological time evolution, it seems natural to have the dimensions start expanding, but we have to explain why only 3 expand, while the other ones remain at the Planck scale. Since we have strings with winding, whose energy is proportional to the size of the compact dimensions ($E = R/\alpha'$), when the dimension grows classical, so do the strings, becoming cosmic strings, and pressing back against the expansion of the dimension. The only way to prevent that is to have annihilation of two closed strings with opposite winding into an unwound string, $w + \bar{w} \leftrightarrow 0$, through a "pair of pants" interaction.

But since the dimension is large and classical, so is the string, so we need to answer whether there is an interaction of *classical* strings, determined solely by geometry. As time evolves, two classical (one dimensional) strings extending in the whole dimension(s) (infinite, if the dimension(s) is/are infinite) oriented arbitrarily and moving in an arbitrary direction, will eventually intersect if the motion is in 3 dimensions, but not if there are 4, since two arbitrary infinite lines form a 3 dimensional space. Of course, there is the case, when the strings are parallel, and never intersect, but that is irrelevant since it has zero measure. Therefore 3 large spatial dimensions is the maximum possible number that allows for annihilation of wound strings of opposite windings, thus avoiding the energy cost that can slow their expansion. In more than 3 spatial dimensions, the windings never annihilate, and the dimensions would stop expanding.

This is then the explanation of 3 large spatial dimensions, even when starting with an arbitrary state of small compact dimensions.

32.3 String Equations of Motion

The idea of the string gas model is very simple: we consider a gas of free (non-interacting) strings, each with Polyakov action. Of course, this is not a very good model: the strings are actually interacting, and non-perturbative, so it is not clear why such a simple model can work so well. On the other hand, in many theories there is some effective description in terms of some free modes, not necessarily the fundamental ones. Perhaps in this case the string gas is made up of effective modes, not the fundamental ones. Or maybe it simply is too simple a model.

For a generic NS-NS background ($g_{\mu\nu}$, $B_{\mu\nu}$, ϕ), the Polyakov action is

$$S_P = -\frac{1}{4\pi\alpha'}\int d^2\sigma \left[\sqrt{-\gamma}\gamma^{ab} g_{\mu\nu}(X)\partial_a X^\mu \partial_b X^\nu + \epsilon^{ab} B_{\mu\nu}(X)\partial_a X^\mu \partial_b X^\nu \right. \\ \left. +\alpha'\sqrt{-\gamma}\phi(X)\mathcal{R}^{(2)}\right]. \tag{32.9}$$

Its equations of motion are:

- the X^μ equation of motion, $\delta S/\delta X^\mu = 0$, giving

$$\partial_a(\sqrt{-\gamma}\gamma^{ab}\partial_b X^\mu) + \Gamma^\mu{}_{\lambda\nu}\sqrt{-\gamma}\gamma^{ab}\partial_a X^\lambda \partial_b X^\nu + \frac{1}{2}H^\mu{}_{\lambda\nu}\epsilon^{ab}\partial_a X^\lambda \partial_b X^\nu = 0\ , \tag{32.10}$$

- the γ^{ab} equation of motion, $\delta S/\delta\gamma^{ab} = 0$,

$$g_{\mu\nu}(X)\left[\partial_a X^\mu \partial_b X^\nu - \frac{1}{2}\gamma_{ab}\gamma^{cd}\partial_c X^\mu \partial_d X^\nu\right] = 0. \tag{32.11}$$

The Polyakov action is classically Weyl invariant, which means that

$$-\frac{2}{\sqrt{-\gamma}}\gamma^{ab}\frac{\delta S_P}{\delta\gamma_{ab}} = T^a{}_a = 0. \tag{32.12}$$

But for local symmetries like Weyl invariance, we need to not have quantum anomalies, i.e., the symmetry must be preserved at the quantum level. Considering a one-loop quantum action on the string worldsheet, i.e., an α' correction, the quantum Weyl (or conformal) invariance condition becomes

$$T^a{}_a = \beta^G_{\mu\nu}\sqrt{-\gamma}\gamma^{ab}\partial_a X^\mu \partial_b X^\nu + \beta^B_{\mu\nu}\epsilon^{ab}\partial_a X^\mu \partial_b X^\nu + \beta^\phi\sqrt{-\gamma}\mathcal{R}^{(2)}\ , \tag{32.13}$$

where

$$\begin{aligned}
\beta^G_{\mu\nu} &= R_{\mu\nu} + 2\nabla_\mu\nabla_\nu\phi - \frac{1}{4}H_{\mu\rho\sigma}H_\nu{}^{\rho\sigma} + \mathcal{O}(\alpha') \\
\beta^B_{\mu\nu} &= \nabla^\rho H_{\rho\mu\nu} - 2\nabla^\rho\phi H_{\rho\mu\nu} + \mathcal{O}(\alpha') \\
\beta^\phi &= \frac{1}{\alpha'}\left(\frac{D-26}{48\pi^2}\right) + 4\nabla_\rho\phi\nabla^\rho\phi - 4\nabla^2\phi - \mathcal{R}^{(2)} + \frac{1}{12}H_{\mu\nu\rho}H^{\mu\nu\rho} + \mathcal{O}(\alpha').
\end{aligned} \tag{32.14}$$

In the last equation, the constant is written for the bosonic string in a noncritical dimension; for the superstring we have a different one.

The equations of motion can be obtained the spacetime effective action

$$S_{\rm ST} = \frac{1}{2\kappa_D^2}\int d^D x\sqrt{-G}e^{-2\phi}\left[R + c + 4(\nabla\phi)^2 - \frac{1}{12}H_{\mu\nu\rho}\right], \tag{32.15}$$

where

$$2\kappa_D^2 = (2\pi\sqrt{\alpha'})^{D-2}\frac{g_s^2}{2\pi}. \tag{32.16}$$

The string background that solves the spacetime equations of motion can be thought of self-consistently as being a condensate of the string modes, but we can still consider other quantum mechanical strings moving in this background.

32.4 String Gas Cosmological Model

The string gas model has T-duality invariance built in, and in the context of the cosmological model, it becomes scale factor duality.

From homogeneity (meaning that the metric components are only functions of time t) and the cosmological ansatz, the solution is

$$\begin{aligned} ds^2 &= -dt^2 + \sum_{i=1}^{d} a_i^2(t)dx_i^2 \\ B_{\mu\nu} &= 0; \quad \phi = \phi(t). \end{aligned} \tag{32.17}$$

Here $d = D - 1$ is the number of space dimensions. We define the log of the scale factors by

$$a_i(t) = e^{\lambda_i(t)}. \tag{32.18}$$

In the case of FLRW cosmology, the scale factors in the 3 noncompact (or rather, large) spatial dimensions are the same, but here we assume a more general ansatz. We also assume that the compact space is a square torus geometry, defined only by the radii (scale factors) of the extra dimensions. The total volume of space (all the dimensions) is

$$V = \prod_{i=1}^{d} e^{\lambda_i}. \tag{32.19}$$

The T-duality invariant dilaton is

$$\tilde{\phi} = \phi - \frac{1}{2}\ln V = \phi - \frac{1}{2}\sum_{i=1}^{d}\lambda_i. \tag{32.20}$$

Then T-duality leaves $\tilde{\phi}$ invariant, while inverting the scale factors,

$$a_i \to \frac{1}{a_i} \Rightarrow \lambda_i \to -\lambda_i . \tag{32.21}$$

To construct the cosmological model, we add a string source S_P to the spacetime action $S_{\rm ST}$, obtaining the modified equations of motion for $g_{\mu\nu}$ and $B_{\mu\nu}$,

$$\begin{aligned}
&\frac{\delta(S_{\rm ST}+S_P)}{\delta g_{\mu\nu}} = 0 \Rightarrow \\
&R_{\mu\nu} + 2\nabla_\mu\nabla_\nu\phi - \frac{1}{4}H_{\mu\rho\sigma}H_\nu{}^{\rho\sigma} = -\frac{\kappa_D^2 e^{2\phi}}{2\pi\alpha'\sqrt{-g}}\int d^2\sigma\sqrt{-\gamma}\gamma^{ab}\partial_a X^\mu \partial_b X^\nu \delta^{(D)}(X - X(\sigma,\tau)) \\
&\frac{\delta(S_{\rm ST}+S_P)}{\delta B_{\mu\nu}} = 0 \Rightarrow \\
&\nabla_\mu(e^{-2\phi}H^{\mu\nu\rho}) = \frac{\kappa_D^2}{\pi\alpha'\sqrt{-g}}\int d^2\sigma\epsilon^{ab}\partial_a X^\mu \partial_b X^\nu \delta^{(D)}(X - X(\sigma,\tau)) \, ,
\end{aligned} \tag{32.22}$$

whereas the ϕ equation is unmodified, since the ϕ term in S_P has the topological $\sqrt{-g}\mathcal{R}^{(2)}$ dependence,

$$4\nabla_\rho\phi\nabla^\rho\phi - 4\nabla^2\phi - \mathcal{R}^{(2)} + \frac{1}{12}H_{\mu\nu\rho}H^{\mu\nu\rho} = 0. \tag{32.23}$$

The energy-momentum tensor of the string source is

$$T^{\mu\nu} = \frac{2}{\sqrt{-g}}\frac{\delta S_P}{\delta g_{\mu\nu}} = -\frac{1}{2\pi\alpha'\sqrt{-g}}\int d^2\sigma\partial_a X^\mu \partial_b X^\nu \delta^{(D)}(X - X(\sigma,\tau)). \tag{32.24}$$

We will use an *adiabatic approximation*, which means that the time variation of the scale factors is slow enough so that we can consider them approximately constant, for the purpose of using flat space formulas for the equations of motion,

$$\partial_+\partial_- X^\mu = 0 \, , \tag{32.25}$$

and for the spectrum. In particular, the left-moving and right-moving momenta for the compact dimensions are as usual

$$\begin{aligned}
p_{R,m} &= \frac{\sqrt{\alpha'}}{R}n_m - \frac{R}{\sqrt{\alpha'}}w_m \\
p_{L,m} &= \frac{\sqrt{\alpha'}}{R}n_m + \frac{R}{\sqrt{\alpha'}}w_m .
\end{aligned} \tag{32.26}$$

In conformal gauge ($\gamma_{ab} = \delta_{ab}$) and using also the light-cone gauge $X^0 = \tau\sqrt{\alpha'}E$ for the residual symmetry (conformal invariance), where $E = p^+$ stands for the energy of the string state, we find

$$\alpha' E^2 = \alpha' \vec{p}^2 + \frac{p_L^2 + p_R^2}{2} + 2(N_L + N_R + a_L + a_R). \tag{32.27}$$

We leave the proof (the same as in flat space) as an exercise. Here $\vec{p}$ is the momentum in the 3 noncompact dimensions, N_L and N_R are occupation numbers for left- and right-movers, and a_L, a_R are the normal ordering constants. In the bosonic string case, $(a_L, a_R) = (-1, -1)$. We also obtain the usual level matching condition

$$p_L^2 - p_R^2 = 4n_i w_i = 4(N_R - N_L - a_R - a_L). \tag{32.28}$$

Eliminating $N_R + a_R$ between the two equations, we find

$$E = \sqrt{\vec{p}^2 + G^{mn}\left(n_m + \frac{w_m}{\alpha'}\right)\left(n_n + \frac{w_n}{\alpha'}\right) + \frac{4}{\alpha'}(N_L + a_L)}\,, \tag{32.29}$$

where G^{mn} is the inverse metric in the compact dimensions.

Replacing this result in the conformal (and lightcone) gauge into the form of the energy-momentum tensor $T_{\mu\nu}$, as well as the solution for X^μ, we obtain

$$T^{00} = \ldots = \frac{E}{\sqrt{-G_{D-1}}}\delta^{(D-1)}(X^i - X^i(\sigma, \tau)). \tag{32.30}$$

We next generalize to N string sources, i.e., a gas made up of N strings. Consider that the gas is split into species s, for instance in the above gauge, parametrized by n_m, w_m, N_L, N_R. Then the energy density is

$$\rho = \sum_s \tilde{n}_s E_s\,, \tag{32.31}$$

where $\tilde{n}_s = n_s/V$ is the number density of species s, and E_s is the energy of a single state of species s. The pressures in directions i are

$$P_i = -\frac{1}{V}\frac{\partial(\rho V)}{\partial \lambda_i}. \tag{32.32}$$

Here V is, as before, the total $d = D - 1$ dimensional volume (in the noncompact *and* the compact directions),

$$V = \prod_{l=1}^{d} e^{\lambda_l}. \tag{32.33}$$

Example

Consider the simplest case, of a gas of bosonic strings, with $a_L = a_R = -1$. Moreover, consider that it is made up only of:

- winding modes, with $w_m \neq 0$, but $n_m = 0$, $N_L = N_R = 1$ (so that $N_L + a_L = N_R + a_R = 0$)
- momentum modes, with $n_m \neq 0$, but with $w_m = 0$, $N_L = N_R = 1$ (so that $N_L + a_L = N_R + a_R = 0$).

Since the radius in direction l is $R_l = e^{\lambda_l}$, and winding modes have energy $n_w^{(l)} R_l = n_w^{(l)} e^{+\lambda_l}$, it follows that the energy density in the case of compactification of all the $d = D - 1$ spatial directions is (now the species s are simply the modes in various compact directions)

$$\rho_w = \sum_{l=1}^{d} \tilde{n}_w^{(l)} e^{\lambda_l(t)} = \frac{1}{V} \sum_{l=1}^{d} n_w^{(l)} e^{\lambda_l(t)} \ , \tag{32.34}$$

and from the general formula for the pressures, we find

$$P_w^{(l)} = -\tilde{n}_w e^{\lambda_l(t)}. \tag{32.35}$$

In the case that all the d directions are identical, we find

$$P_w = -\frac{1}{d}\rho_w. \tag{32.36}$$

For the momentum modes, the energy is $n_m^{(l)}/R_l = n_w^{(l)} e^{-\lambda_l}$, so that the energy density is

$$\rho_m = \sum_{l=1}^{d} \tilde{n}_m^{(l)} e^{-\lambda_l(t)} = \frac{1}{V} \sum_{l=1}^{d} n_m^{(l)} e^{-\lambda_l(t)} \ , \tag{32.37}$$

the pressures are now

$$P_m^{(l)} = \tilde{n}_m e^{-\lambda_l(t)} \ , \tag{32.38}$$

and in the case that all d directions are identical, we find

$$P_m = +\frac{1}{d}\rho_m. \tag{32.39}$$

The results are consistent with the fact that winding modes provide a kind of zero point energy, leading to negative pressure, whereas momentum modes give the usual positive pressure.

32.5 Cosmological Early Time Evolution

We can now consider a simplified model of *cosmological* early time evolution, in which the noncompact dimensions take the FLRW form, i.e., with equal scale factors, $\lambda_1 = \lambda_2 = \lambda_3 \equiv \lambda$, whereas for simplicity we consider all the compact scale factors to be identical, $\lambda_4 = \cdots = \lambda_d = \nu$. We consider also a solution with a constant NS-NS field strength H in the noncompact spatial directions H_{123}, i.e.,

$$\begin{aligned} ds^2 &= -dt^2 + e^{2\lambda(t)} d\vec{x}^2 + e^{2\nu(t)} d\vec{y}^2 \\ H_3 &= h dx^1 \wedge dx^2 \wedge dx^3 \\ \phi &= \phi(t). \end{aligned} \tag{32.40}$$

Though it might seem that a constant field with indices, that picks up a certain direction or directions as special breaks Lorentz invariance, in this case it doesn't break more than what the FLRW solution already does. Indeed, as we already mentioned when talking about the CMBR and the fact that we can measure the velocity of the Earth with respect to it, since we are in general relativity with a non-Minkowski metric, we don't have a (global) Lorentz invariance anymore, only a local Lorentz invariance in an infinitesimal neighbourhood of a point. The (global) cosmological FLRW solution picks out a special (global) frame, that defines the cosmological time. This is equivalent to picking out a special vector A_0, or equivalently, the Poincaré dual spatial 3-tensor H_{123}.

Adding the string sources, as we have seen, adds on the right hand side of the $g^{\mu\nu}$ and ϕ equations of motion terms involving $T_{\mu\nu}$, as

$$\begin{aligned} &R_{\mu\nu} + 2\nabla_\mu \nabla_\nu \phi - \frac{1}{4} H_{\mu\rho\sigma} H_\nu{}^{\rho\sigma} = 2\kappa^2 e^{2\phi} T_{\mu\nu} = 16\pi G_{10} e^{2\phi} T_{\mu\nu} \\ &4\nabla_\rho \phi \nabla^\rho \phi - 4\nabla^2 \phi - \mathcal{R}^{(2)} + \frac{1}{12} H_{\mu\nu\rho} H^{\mu\nu\rho} = 2\kappa^2 e^{2\phi} T^\mu{}_\mu = 16\pi G_{10} e^{2\phi} T^\mu{}_\mu. \end{aligned} \tag{32.41}$$

The $H_{\mu\nu\rho}$ equation of motion is automatically satisfied by the ansatz, as one can check.

With this simplified ansatz, the total spacetime volume $V = e^{3\lambda} e^{6\nu}$, so the energy of the string gas is

$$E = \rho V = \rho e^{3\lambda + 6\nu} \,, \tag{32.42}$$

and the pressures times volume are

$$\bar{P}_i \equiv P_i V = P_i e^{3\lambda + 6\nu} \,, \tag{32.43}$$

and they split into P_λ (usual pressures, times volume, in the noncompact directions), and P_ν (in the compact dimensions).

The Einstein ($g_{\mu\nu}$) and dilaton (ϕ) equations of motion become on our ansatz

$$\begin{aligned} & c - 3\dot{\lambda} - 6\dot{\nu}^2 + 4\dot{\tilde{\phi}}^2 - \frac{h^2}{2}e^{-6\lambda} = e^{2\tilde{\phi}}E \\ & \ddot{\lambda} - 2\dot{\tilde{\phi}}\dot{\lambda} - \frac{h^2}{2}e^{-6\lambda} = \frac{1}{2}e^{2\tilde{\phi}}\bar{P}_\lambda \\ & \ddot{\nu} - 2\dot{\tilde{\phi}}\dot{\nu} = \frac{1}{2}e^{2\tilde{\phi}}\bar{P}_\nu \\ & 2\ddot{\tilde{\phi}} - 3\dot{\lambda}^2 - 6\dot{\nu}^2 = \frac{1}{2}e^{2\tilde{\phi}}E. \end{aligned} \tag{32.44}$$

We also need to impose the conservation equation for the energy-momentum tensor, $\partial^\mu T_{\mu\nu} = 0$, which becomes

$$\dot{E} + 3\dot{\lambda}\bar{P}_\lambda + 6\dot{\nu}\bar{P}_\nu = 0. \tag{32.45}$$

These equations, with ρ, P_λ, P_ν given by the gas of string momentum and winding modes, cannot be solved analytically, but we can make simulations of the time evolution.

One finds a period of "cosmological loitering", during which $\dot{\lambda} = 0$, $\dot{\tilde{\phi}} = 0$, i.e., our 3 spatial dimensions don't evolve. (in fact, in the Einstein frame, they contract) This lasts until all winding modes annihilate. After its end, our 3 spatial dimensions can grow large. The existence of this period solves the horizon problem in a different way than in inflation. Rather than having a small patch blow exponentially way outside the horizon, the size remains constant as the Hubble size evolves, but the effect is the same. Also the isotropy (if not the homogeneity) problem is solved, as initial differences between the various directions get smoothed over.

32.6 Moduli Stabilization

All the "radions", the radii (or sizes) of the extra dimensions, will be stabilized by the string gas model. Going back to the general ansatz for the compact dimensions, i.e.,

$$e^{2\nu}d\vec{y}^2 \to \sum_{\alpha=1}^{6} b_\alpha^2(t)dy_\alpha^2 \,, \tag{32.46}$$

one finds the equation of motion for them

$$\ddot{b}_\alpha + \left(3H + \sum_{\beta=1,\beta\neq\alpha}^{6}\frac{\dot{b}_\beta}{b_\beta}\right)\dot{b}_\alpha = \sum_{m,n} 8\pi G_N \frac{\mu_{m,n}}{\epsilon_{m,n}} S \,, \tag{32.47}$$

where $\mu_{m,n}$ is the number density of string states with momentum n nad winding m, $\epsilon_{m,n}$ is the energy of this state, and

$$S = \sum_{\alpha} \left(\frac{m_\alpha}{b_\alpha} \right)^2 - \sum_{\alpha} n_\alpha^2 b_\alpha^2 + \frac{2}{D-1} \left[\vec{n} \cdot \vec{n} + \vec{n} \cdot \vec{m} + 2(N-1) \right]. \quad (32.48)$$

The last term will vanish at the self-dual (T-duality invariant) radius, i.e., it will be stabilized (since it minimizes the right hand side of the equation of motion, the "potential").

Important Concepts to Remember

- For an inflation that lasts more than 70 e-folds, we find that current Hubble scales were initially sub-Planckian, hence it makes no sense to treat them in effective field theory (inflation).
- The Hagedorn temperature $T_H = 1/(4\pi\sqrt{\alpha'})$ is the maximum temperature of a gas of strings, after which the partition function diverges due to the exponential number density of states.
- The existence of the Hagedorn temperature is related to T-duality invariance. The calculation of the density of states for a fully compactified theory (the only case that gives positive specific heat $C > 0$) gives the exponential behaviour.
- From the initially all compactified dimensions, only 3 expand and become large, since only for those can strings find each other, interact, and annihilate their winding (which otherwise keeps the dimensions small).
- The string equations of motion are obtained from quantum conformal invariance of the string worldsheet, and can be derived from a spacetime effective action.
- Considering a gas of strings, with momentum modes and winding modes, the winding modes have equation of state $P = -\rho/d$ and the momentum modes have equation of state $P = +\rho/d$.
- Early time cosmology has a "cosmological loitering" phase, in which the 3 spatial dimensions don't evolve (in fact, contract in the Einstein frame), and winding modes annihilate. It solves the horizon and isotropy problem differently than in inflation.
- Moduli get stabilized during the time evolution.

Further reading: The original paper of Brandenberger and Vafa is [57]. Good reviews for string gas cosmology are [58, 59] and the recent one [60].

Exercises

(1) Prove that for $N_e > 70$-folds of inflation current cosmological scales ($\sim H_0^{-1}$) were $l_{\rm Pl}$ at initial time.

(2) Show that in conformal gauge, with $X^\rho = \tau\sqrt{\alpha'}E$, we obtain

$$\alpha' E^2 = \alpha' \vec{p}^2 + \frac{p_L^2 + p_R^2}{2} + 2(N_L + N_R + a_L + a_R). \quad (32.49)$$

(3) Show that the conservation equation $\partial^\mu T_{\mu\nu} = 0$ becomes

$$\dot{E} + 3\dot{\lambda}\bar{P}_\lambda + 6\dot{\nu}\bar{P}_\nu = 0. \tag{32.50}$$

(4) Show that the equations of motion of the spacetime action plus the string source, $S_{\rm ST} + S_P$, are given by (32.22).

Chapter 33
String Gas and Brane Gas Developments

In this chapter we will first describe a simple generalization of the previously defined string gas cosmology, namely brane gas; and then we will calculate the spectrum of scalar and tensor perturbations in string gas cosmology.

33.1 Brane Gas

The idea of the brane gas is obvious from its name: replace the gas of strings from the previous chapter with a gas of p-branes.

The p-brane action is taken to be a DBI action for interaction with a NS-NS background, and with an electromagnetic field on its worldvolume, so

$$S_{\text{DBI,p}} = -T_p \int d^{p+1}\xi e^{-\phi}\sqrt{-\det(\tilde{g}_{ab} + \tilde{B}_{ab} + 2\pi\alpha' F_{ab})}\,, \tag{33.1}$$

where $\tilde{g}_{ab}$ and $\tilde{B}_{ab}$ are induced NS-NS metric and B-fields on the $p+1$ dimensional worldvolume, and F_{ab} is the field strength of the electromagnetic field on the worldvolume,

$$\begin{aligned} \tilde{g}_{ab} &= g_{\mu\nu}\partial_a X^\mu \partial_b X^\nu \\ \tilde{B}_{ab} &= B_{\mu\nu}\partial_a X^\mu \partial_b X^\nu . \end{aligned} \tag{33.2}$$

In a static gauge, $X^a = \xi^a$, $X^i = \phi^i$, we find

$$\begin{aligned} \tilde{g}_{ab} &= g_{ab} + g_{ij}\partial_a\phi^i\partial_b\phi^j + g_{ib}\partial_a\phi^i + g_{ia}\partial_b\phi^i \\ \tilde{B}_{ab} &= B_{ab} + B_{ij}\partial_a\phi^i\partial_b\phi^j + B_{i[a}\partial_{b]}\phi^i . \end{aligned} \tag{33.3}$$

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_33

The tension of the p-brane is

$$T_p = \frac{1}{(2\pi\sqrt{\alpha'})^{p+1}} \frac{\pi}{g_s} \tag{33.4}$$

We also consider a cosmological background in conformal time, where for simplicity we consider the same scale factor for all directions,

$$\begin{aligned} ds^2 &= a^2(\eta)[-d\eta^2 + d\vec{x}^2] \\ B_{\mu\nu} &= 0. \end{aligned} \tag{33.5}$$

Substituting it into the DBI action, and expanding in the number of derivatives, we find

$$S_{\text{DBI,p}} \simeq -T_p \int d^{p+1}\xi a(\eta)^{p+1} e^{-\phi} \left[1 + \frac{1}{2}(\partial_a \phi)^2 + \pi^2 \alpha'^2 a^{-4} F_{ab}^2 + \cdots \right]. \tag{33.6}$$

We can now calculate the energy-momentum tensor T^{ab} for the analog of winding modes, which are now "wrapped" modes of the brane (all p directions wrap some of the -compact- spatial directions). From it, we can calculate the energy density ρ and the pressure P, obtaining

$$P_i = w_i \rho; \quad w_p = -\frac{p}{d}. \tag{33.7}$$

Here d is the total number of space dimensions. For a 3-brane in $3+1$ dimensional space, we would obtain $w = -1$, since this would then truly be a cosmological constant. On the other hand, for momentum modes, we have as usual $w = +1/d$, whereas in general normal matter has $w \in [0, 1]$.

The equations of motion for the cosmology are obtain from the equations of the previous chapter (32.44), considered for $c = 0$ (i.e., in the critical dimension, $D = 26$ for the bosonic string and $D = 10$ for the superstring), $h = 0$ (since we now consider only $B_{\mu\nu} = 0$), and $\lambda = \nu$ since we now consider the same scale factor for all dimensions, which is certainly true during the first, *Hagedorn*, or cosmological loitering, phase of the cosmological evolution, before our 3 spatial dimensions start expanding. Indeed, in the following we will only consider this Hagedorn phase. The equations are then

$$\begin{aligned} -d\dot{\lambda} + 4\dot{\tilde{\phi}}^2 &= e^{2\tilde{\phi}} E \\ \ddot{\lambda} - 2\dot{\tilde{\phi}}\dot{\lambda} &= \frac{1}{2} e^{2\tilde{\phi}} \bar{P} \\ 2\ddot{\tilde{\phi}} - d\dot{\lambda}^2 &= \frac{1}{2} e^{2\tilde{\phi}} E. \end{aligned} \tag{33.8}$$

Here as usual $\tilde{\phi} = \phi - \frac{d}{2}\lambda$ is the T-duality invariant dilaton. The energy and pressure follow from the previous brane action. Indeed, we can say that, since the energy is proportional to the brane tension times the spatial volume,

$$E \propto T_p[a(\eta)]^p \ , \tag{33.9}$$

the effective potential in 10 dimensional string frame is

$$V_{\text{eff}}(\lambda) = \beta_p e^{2\tilde{\phi}} e^{p\lambda}. \tag{33.10}$$

Here $e^{p\lambda} = a(\eta)^p$ is the spatial volume of the p-brane, $\beta_p \propto T_p$. In more generality, if our 3 dimensions have scale factor a and the other d have scale factor b (valid after the Hagedorn phase), the effective potential in string frame would be

$$V_s^{(4+d)} = \frac{N}{(2\pi\sqrt{\alpha'})^4} \frac{b^k}{a^3 b^d} \ , \tag{33.11}$$

where the $1/(2\pi\sqrt{\alpha'})^4 = T_3$ is a 3-brane tension, N is the number of branes, $a^3 b^d$ is the total volume of space, and $k = 1$ for the p-branes and $k = -1$ for momentum modes is the dependence coming either from the p-dimensional volume of the branes wrapped over the compact space (so that $p \leq d$) or from the $1/R = 1/b$ momentum of the momentum modes.

The importance of the p-brane gas model comes from a generalization of the argument about the string gas selecting 3 large dimensions. For two generic classical p-branes to intersect as time evolves, we need a large $2p + 1$ dimensional space, since two generic p-branes form a $2p + 1$ dimensional space. That means that, through the interaction, we can eventually annihilate all p-branes of opposite wrapping number only if the dimension of space is $2p + 1$ or smaller. But higher p branes annihilate slower than lower p ones (there's a larger phase space to be covered), which means that we create a hierarchy of dimensions.

Then, for instance, in the presence of strings and 2-branes (which annihilate in 5+1 dimensions), we will create 3 very large (almost infinite) dimensions (because of the strings), but also two larger dimensions (because of the 2-branes), ending up with a *large extra dimensions* scenario. This is the main interest of the brane gas: the possibility of hierarchies for dimensions, allowing for a large extra dimensions scenario.

33.2 Cosmological Perturbations Generalities

The cosmological evolution we will be interested in happens in a *Hagedorn phase*. Like we said in the previous chapter, during this phase (also called cosmological loitering phase), in Einstein frame the horizon scale $H^{-1}(t)$ *shrinks*, solving the

horizon problem in the opposite way to inflation: there the scales blow up exponentially, and H^{-1} is fixed, here scales are evolving slowly, but the horizon shrinks. It in fact shrinks until it reaches Hagedorn time t_R, where it is minimum,

$$H^{-1}(t_R) \sim \frac{M_{\rm Pl}}{T_H^2} \ , \tag{33.12}$$

and then, for $t > t_R$, when many scales have gone already outside the horizon, we start standard cosmology, and H^{-1} starts increasing again. Perturbations observed in the CMBR will be created close to this minimum, so this is what we will try to calculate.

We therefore move on to the calculation of cosmological perturbations, which will leave an imprint in the CMBR. We want to calculate the scalar and tensor perturbations, using a starting point similar to that for inflation: The most general perturbation of the metric with scalar and tensor components, in a Newtonian gauge with $\Phi = \Psi$ and in conformal time, is

$$ds^2 = a^2(\eta) \left\{ -(1 + 2\Phi(x, \eta))d\eta^2 + [(1 - 2\Phi(x, \eta))\delta_{ij} + h_{ij}]dx^i dx^j \right\} . \tag{33.13}$$

Here Φ will generate scalar perturbations, and h_{ij} tensor perturbations.

For the Newton potential Φ, we have the Newton equation relating it to a small density fluctuation acting as a source,

$$\nabla^2 \Phi = 4\pi G_N \delta\rho. \tag{33.14}$$

In momentum space, we have

$$-k^2 \Phi(k) = 4\pi G_N \delta\rho(k) = 4\pi G_N \delta T^0{}_0(k). \tag{33.15}$$

For the tensor perturbation h_{ij}, for $i \neq j$, we have similarly

$$\nabla^2 h_{ij} = 4\pi G_N \delta T^i{}_j \ , \tag{33.16}$$

or in momentum space

$$-k^2 h_{ij}(k) = 4\pi G_N \delta T^i{}_j. \tag{33.17}$$

Substituting in the correlators that we want to calculate, we obtain (for $i \neq j$)

$$\begin{aligned} \langle |\Phi(k)|^2 \rangle &= \frac{(4\pi G_N)^2}{k^4} \langle \delta T^0{}_0(k) \delta T^0{}_0(k) \rangle \\ \langle |h(k)|^2 \rangle &= \frac{(4\pi G_N)^2}{k^4} \langle \delta T^i{}_j(k) \delta T^i{}_j(k) \rangle. \end{aligned} \tag{33.18}$$

33.3 Correlators from Thermodynamics

We now express the energy-momentum tensor correlators in terms of thermodynamical quantities, which will be later calculated for the string gas.

For a general system with a metric component g_{00}, the partition function Z, free energy F and Hamiltonian H, for states s, are related by

$$e^{-\beta F} = Z = \sum_s e^{-\beta\sqrt{-g_{00}}H(s)}. \tag{33.19}$$

For the strings, with a nontrivial g_{00}, but in a static configuration, so the Lagrangian equals minus the Hamiltonian, $\mathcal{L} = -H$, we have the action (written in terms of the proper time $d\tau = dt\sqrt{-g_{00}}$)

$$S = \int dt\sqrt{-g_{00}}H. \tag{33.20}$$

In a path integral formulation in Euclidean space, $\int dt = \beta = 1/T$.

But since the energy-momentum tensor is

$$T_{\mu\nu} = \frac{2}{\sqrt{-g}}\frac{\delta S}{\delta g^{\mu\nu}} \Rightarrow T^{\mu}{}_{\nu} = \frac{2g^{\mu\lambda}}{\sqrt{-g}}\frac{\delta S}{\delta g^{\lambda\nu}}\ , \tag{33.21}$$

and the variational derivative δ means normal derivative d, when we drop the time integral $\int dt$, we obtain in a state s

$$T^{\mu}{}_{\nu}(s) = \frac{2g^{\mu\lambda}}{\sqrt{-g}}\frac{d}{dg^{\lambda\nu}}[-\sqrt{-g_{00}}H(s)]. \tag{33.22}$$

But a general ensemble average gives

$$\langle M\rangle \equiv \frac{1}{Z}\sum_s e^{-\beta\sqrt{-g_{00}}H(s)}M\ , \tag{33.23}$$

so applying it for $T^{\mu}{}_{\nu}$ gives

$$\langle T^{\mu}{}_{\nu}\rangle = \frac{2g^{\mu\lambda}}{\sqrt{-g}}\frac{\delta}{\delta g^{\lambda\nu}}\ln Z\ , \tag{33.24}$$

since

$$\delta e^{-\beta\sqrt{-g_{00}}H(s)} = e^{-\beta\sqrt{-g_{00}}H(s)}\delta[-\beta\sqrt{-g_{00}}H(s)] = e^{-\beta\sqrt{-g_{00}}H(s)}d[-\sqrt{-g_{00}}H(s)]. \tag{33.25}$$

Here we have used that $\beta = \int dt$ and the system is time-independent.

Now considering a second derivative will give the correlator of the fluctuations of the energy-momentum tensor,

$$\begin{aligned}\langle \delta T^{\mu}{}_{\nu} \delta T^{\sigma}{}_{\rho} \rangle &= \langle T^{\mu}{}_{\nu} T^{\sigma}{}_{\rho} \rangle - \langle T^{\mu}{}_{\nu} \rangle \langle T^{\sigma}{}_{\rho} \rangle \\ &= \frac{2g^{\mu\alpha}}{\sqrt{-g}} \frac{\delta}{\delta g^{\alpha\nu}} \left(\frac{g^{\sigma\lambda}}{\sqrt{-g}} \frac{\delta}{\delta g^{\lambda\rho}} \ln Z \right) + \frac{2g^{\sigma\alpha}}{\sqrt{-g}} \frac{\delta}{\delta g^{\alpha\rho}} \left(\frac{g^{\mu\lambda}}{\sqrt{-g}} \frac{\delta}{\delta g^{\lambda\nu}} \ln Z \right) .\end{aligned} \tag{33.26}$$

Applying this for $\mu = \nu = \rho = \sigma = 0$, we obtain

$$\begin{aligned}\langle \delta T^{0}{}_{0} \delta T^{0}{}_{0} \rangle &= \langle (\delta\rho)^2 \rangle \\ &= \frac{4g^{00}}{\sqrt{-g}} \frac{\delta}{\delta g^{00}} \left(\frac{g^{00}}{\sqrt{-g}} \frac{\delta}{\delta g^{00}} \ln Z \right) .\end{aligned} \tag{33.27}$$

Considering a 3 dimensional spatial metric $g_{ij} \simeq R^2 \delta_{ij}$, we obtain

$$\langle (\delta\rho)^2 \rangle = \frac{1}{R^6} \left(\frac{2g_{00}}{\sqrt{-g_{00}}} \frac{\delta}{\delta g_{00}} \right)^2 \ln Z = \frac{1}{R^6} \left(\frac{\delta}{\delta \sqrt{-g_{00}}} \right)^2 \ln Z . \tag{33.28}$$

But since $\ln Z = -\beta F$, and, when acting on Z, $\delta/\delta\sqrt{-g_{00}} = d/(\sqrt{-g_{00}} d\beta)$, we obtain

$$\langle (\delta\rho)^2 \rangle = -\frac{1}{R^6(-g_{00})} \frac{d^2}{d\beta^2} (\beta F) = -\frac{1}{R^6(-g_{00})} \frac{\partial}{\partial\beta} \left(F + \beta \frac{\partial}{\partial\beta} F \right) . \tag{33.29}$$

From the thermodynamic relations $F = E - TS$ and $dF = -SdT - PdV$, with $V = R^3$, we find

$$E = F + TS = F + \beta \left(\frac{\partial F}{\partial \beta} \right) , \tag{33.30}$$

which means that

$$T^2 C_V = T^2 \left(\frac{\partial E}{\partial T} \right)_V = -\left(\frac{\partial E}{\partial \beta} \right)_V = -\frac{\partial}{\partial\beta} \left(F + \beta \left(\frac{\partial F}{\partial \beta} \right)_V \right)_V . \tag{33.31}$$

Finally, we obtain

$$\langle (\delta\rho)^2 \rangle = \frac{T^2}{R^6} C_V . \tag{33.32}$$

Next we apply the same methods for the spatial components,

$$\begin{aligned}\langle \delta T^{i}{}_{j} \delta T^{i}{}_{j} \rangle &= \langle T^{i}{}_{j} T^{i}{}_{j} \rangle - \langle T^{i}{}_{j} \rangle \langle T^{i}{}_{j} \rangle \\ &= \frac{4g^{ik}}{\sqrt{-g}} \frac{\delta}{\delta g^{kj}} \left(\frac{g^{il}}{\sqrt{-g}} \frac{\delta}{\delta g^{lj}} \ln Z \right) ,\end{aligned} \tag{33.33}$$

with $i \neq j$. With $g_{ij} = R^2(\delta_{ij} + h_{ij}) \simeq R^2\delta_{ij}$, we obtain

$$\langle \delta T^i{}_j \delta T^i{}_j \rangle = \frac{4\delta R^2}{\sqrt{-g_{00}}R^3} \frac{\delta}{\delta R^2} \left(\frac{R^2}{\sqrt{-g_{00}}R^3} \frac{\delta}{\delta R^2} \right) \ln Z = \left(\frac{1}{\sqrt{-g_{00}}R^3} \frac{\delta}{\delta \ln R} \right)^2 \ln Z. \tag{33.34}$$

But since $\delta/\delta R$, when acting on Z, removes also an $\int dt = \beta$, so $\delta/\delta R = (1/\beta)d/dR$, and $\ln Z = -\beta F$, we obtain

$$\langle \delta T^i{}_j \delta T^i{}_j \rangle = \frac{1}{-g_{00}\beta} \frac{1}{R^3} \frac{\partial}{\partial \ln R} \left(-\frac{1}{R^3} \frac{\partial F}{\partial \ln R} \right). \tag{33.35}$$

Then, since $dE = TdS - PdV$ and $dF = -SdT - PdV$ and $V = R^3$, we have

$$p = T \left(\frac{\partial S}{\partial V} \right)_E = -\left(\frac{\partial F}{\partial V} \right)_T = -\frac{1}{3R^3} \left(\frac{\partial F}{\partial \ln R} \right)_T \tag{33.36}$$

so that finally

$$\langle \delta T^i{}_j \delta T^i{}_j \rangle = \frac{3}{(-g_{00})\beta R^2} \left(\frac{\partial p}{\partial R} \right)_T. \tag{33.37}$$

33.4 Thermodynamics for String Gas

With the general thermodynamics formulas, we could calculate the fluctuation correlator for any quantum gas. But in particular, we want to apply for the gas of strings. We need to calculate the total number of states of given energy E and volume R^3, $\Omega(E, R)$, and from it the entropy

$$S(E, R) = \ln \Omega(E, R). \tag{33.38}$$

A string calculation finds

$$\Omega(E, R) \simeq \beta_H e^{\beta_H E + n_H V} [1 + \delta\Omega_{(1)}], \tag{33.39}$$

where $\beta_H = 1/T_H$ with T_H the Hagedorn temperature, $n_H \sim l_s^{-3}$ is the Hagedorn number density and $\delta\Omega_{(1)}$, with $-\delta\Omega_{(1)} \ll 1$, comes from a density of states, and is found to be

$$\delta\Omega_{(1)} \simeq -\frac{(\beta_H E)^5}{5!} e^{-(\beta_H - \beta_1)(E - \rho_H V)}, \tag{33.40}$$

where $\rho_H \sim l_s^{-4}$ is a Hagedorn density, and β_1 is the closest singularity to β_H, which obeys

$$\beta_H - \beta_1 \sim \frac{l_s^3}{R^2}; \qquad R \gg l_s$$
$$\sim \frac{R^2}{l_s}; \qquad R \ll l_s. \tag{33.41}$$

We then obtain

$$S(E, R) \simeq \beta_H E + n_H V + \ln[1 + \delta\Omega_{(1)}] \,, \tag{33.42}$$

and, since $\frac{1}{T} = \left(\frac{\partial S}{\partial E}\right)_V$, we find

$$T(E, R) = \left[\left(\frac{\partial S}{\partial E}\right)_V\right]^{-1} \simeq \left[\beta_H + \left(\frac{\partial \delta\Omega_{(1)}}{\partial E}\right)_V\right]^{-1} \simeq T_H \left[1 - \frac{1}{\beta_H}\left(\frac{\partial \Omega_{(1)}}{\partial E}\right)_V\right]$$
$$\simeq T_H \left[1 + \frac{\beta_H - \beta_1}{\beta_H}\delta\Omega_{(1)}\right]. \tag{33.43}$$

Inverting this relation, and using $\beta_H - \beta_1 \simeq l_s^3/R^2$, since we are interested in the regime $R \gg l_s$, we obtain

$$l_s^3 \delta\Omega_{(1)} \simeq -\frac{R^2}{T_H}\left(1 - \frac{T}{T_H}\right). \tag{33.44}$$

Then inverting the relation (33.40) to find E as a function of $\Omega_{(1)}$, we obtain

$$E \simeq \frac{1}{\beta_H - \beta_1} \ln \frac{1}{-\delta\Omega_{(1)}} \simeq \frac{R^2}{l_s^3} \ln \left[\frac{l_s^3 T}{R^2\left(1 - \frac{T}{T_H}\right)}\right]. \tag{33.45}$$

Then finally

$$C_V = \left(\frac{\partial E}{\partial T}\right)_V \simeq \frac{R^2}{l_s^3}\frac{\partial}{\partial T} \ln \frac{T}{1 - \frac{T}{T_H}} = \frac{R^2/l_s^3}{T\left(1 - \frac{T}{T_H}\right)}. \tag{33.46}$$

For the pressure, the formula for the entropy (33.42) gives

$$p = T\left(\frac{\partial S}{\partial V}\right)_E \simeq n_H T + T\left(\frac{\partial}{\partial V}\delta\Omega_{(1)}\right)_E. \tag{33.47}$$

But

$$\left(\frac{\partial \delta\Omega_{(1)}}{\partial V}\right)_E = \left[-\frac{\partial(\beta_H - \beta_1)}{\partial V}(E - \rho_H V) + \rho_H(\beta_H - \beta_1)\right]\delta\Omega_{(1)}. \tag{33.48}$$

Using $\beta_H - \beta_1 = l_s^3/R^2$, $\rho_H = l_s^{-4}$, $V = R^3$, E from (33.45) and $\delta\Omega_{(1)}$ from (33.44), we find after some algebra

$$p \simeq n_H T - \frac{T}{T_H}\frac{1-\frac{T}{T_H}}{l_s^3}\left\{\frac{5}{3l_s} + \frac{2}{3R}\ln\left[\frac{l_s^3 T}{R^2\left(1-\frac{T}{T_H}\right)}\right]\right\}. \tag{33.49}$$

If we are very close to the Hagedorn temperature, $T \simeq T_H$, we can approximate further

$$p \simeq n_H T_H - \frac{2}{3}\frac{1-\frac{T}{T_H}}{l_s^3 R}\ln\left[\frac{l_s^3 T}{R^2\left(1-\frac{T}{T_H}\right)}\right] \tag{33.50}$$

Moreover, we can calculate the derivative

$$\left(\frac{\partial p}{\partial R}\right)_T \simeq \frac{T}{T_H}\frac{1-\frac{T}{T_H}}{l_s^3 R^2}\left(\frac{4}{3} + \frac{2}{3}\ln\left[\frac{l_s^3 T}{R^2\left(1-\frac{T}{T_H}\right)}\right]\right). \tag{33.51}$$

Finally then, we can calculate the energy-momentum tensor correlators,

$$\langle(\delta\rho)^2\rangle = \frac{T^2}{R^6}C_V \simeq \frac{l_s^{-3}}{R^4}\frac{T}{\left(1-\frac{T}{T_H}\right)} \tag{33.52}$$

and, for $i \neq j$,

$$\langle \delta T^i{}_j \delta T^i{}_j\rangle = \frac{3T}{R^2}\left(\frac{\partial p}{\partial R}\right)_T \simeq \frac{3T^2}{T_H}\frac{1-\frac{T}{T_H}}{l_s^3 R^4}\left(\frac{4}{3} + \frac{2}{3}\ln\left[\frac{l_s^3 T}{R^2\left(1-\frac{T}{T_H}\right)}\right]\right). \tag{33.53}$$

For $T \simeq T_H \simeq l_s^{-1}$, we can approximate further,

$$\langle \delta T^i{}_j \delta T^i{}_j\rangle \simeq \frac{2T\left(1-\frac{T}{T_H}\right)}{l_s^3 R^4}\ln\left[\frac{l_s^2}{R^2\left(1-\frac{T}{T_H}\right)}\right]. \tag{33.54}$$

33.5 String Gas Power Spectra and Tilts

We can finally relate to CMBR physics. The correlators of $\Phi(k)$ and $|h(k)|$ give rise to scalar and tensor power spectra in the CMBR. It is assumed that the fluctuations in the CMBR are close to the end of the Hagedorn phase, and the beginning of standard

cosmology, like in the case of inflation (where the fluctuations are computed towards the end of inflation).

Scalar Power Spectrum

As we saw in part I of the book, the scalar power spectrum in the CMBR and its tilt n_s is defined by

$$P_\Phi(k) = \frac{k^3}{2\pi^2}|\Phi(k)|^2 \propto k^{n_s-1} \ , \tag{33.55}$$

so

$$n_s - 1 = \frac{d}{d\ln k} \ln P_\Phi(k). \tag{33.56}$$

But $\langle|\Phi(k)|^2\rangle$ was related to $\langle(\delta\rho)^2\rangle$ in (33.18), which gives

$$P_\Phi(k) = \frac{8G_N^2}{k}\langle(\delta\rho(k))^2\rangle. \tag{33.57}$$

Note however that we need a density fluctuation in momentum space, $\delta\rho(k)$, whereas the correlators we have calculated were in position space. We can relate them by considering a sphere of radius $R = 1/k$. Then $\delta\rho = \delta M/R^3 = k^3\delta M$, whereas in momentum space (because of the Fourier transform)

$$\langle(\delta\rho(k))^2\rangle = \frac{1}{V}\langle(\delta M)^2\rangle = k^3\langle(\delta M)^2\rangle = \frac{1}{k^3}\langle(\delta\rho)^2\rangle. \tag{33.58}$$

Then, from the density correlator (33.52), we find

$$P_\Phi(k) = \frac{8G_N}{l_s^3}\frac{T(k)}{1-\frac{T(k)}{T_H}}. \tag{33.59}$$

We can rewrite this, using $T_H \sim l_s^{-1}$, as

$$P_\Phi(k) \sim \left(\frac{l_{\rm Pl}}{l_s}\right)^4 \frac{T(k)}{T_H}\frac{1}{1-\frac{T(k)}{T_H}} \tag{33.60}$$

The first observation ia that the spectrum is almost scale invariant, $n_s - 1 \simeq 0$, since there is no explicit k dependence, and only a very small scale implicit dependence in the temperature $T(k)$. That is due to the fact that, as in inflation, we calculate the fluctuation when the scale exits the horizon (since it is frozen in there, and then goes back inside the horizon only now), at $t_i(k)$, just before the minimum horizon size time t_R.

But moreover, we have a slight red tilt, as observed experimentally, since, like in inflation, larger scales (larger wavelengths λ, thus smaller k) exit the horizon sooner (at smaller $t_i(k)$), when the temperature $T(k)$ is higher: the temperature slightly

decreases with time at the end of the Hagedorn phase, while remaining smaller than the maximum one, T_H. That means that $dT(k)/dk < 0$.

Specifically, we obtain

$$n_s - 1 = \frac{d \ln P_\Phi}{d \ln k} = \left(\frac{1}{1 - \frac{T(k)}{T_H}} + 1 \right) k \frac{d}{dk} \frac{T(k)}{T_H} \simeq \frac{1}{1 - \frac{T(k)}{T_H}} k \frac{d}{dk} \frac{T(k)}{T_H} < 0. \tag{33.61}$$

It is however impossible to calculate analytically dT/dk, so that is the most we can say: $n_s - 1$ can be small and negative, as observed experimentally.

Tensor Power Spectrum

Similarly, we define for the tensor perturbation

$$P_h(k) = \frac{k^3}{2\pi^2} |h(k)|^2 \propto k^{n_T - 1}. \tag{33.62}$$

We have defined the power with a -1, even though sometimes it is defined without it, in order to have symmetry with the scalar spectrum.

Then in the same way as for the scalar case, we find

$$P_h(k) = \frac{8 G_N^2}{k} \langle \delta T^i{}_j(k) \delta T^i{}_j(k) \rangle \ , \tag{33.63}$$

but we need to relate to the previously calculated one by considering a sphere of radius $R = 1/k$, so

$$P_h(k) = \frac{8 G_N^2}{k^4} \langle \delta T^i{}_j(R) \delta T^i{}_j(R) \rangle \ , \tag{33.64}$$

and, using (33.53), we find

$$P_h(k) \simeq \frac{16 G_N^2}{l_s^3} \frac{T^2 \left(1 - \frac{T}{T_H}\right)}{T_H} \ln \left[\frac{l_s^2}{R^2 \left(1 - \frac{T}{T_H}\right)} \right] , \tag{33.65}$$

that becomes

$$P_h(k) \simeq \left(\frac{l_{\rm Pl}}{l_s} \right)^4 \frac{T^2}{T_H^2} \left(1 - \frac{T}{T_H}\right) \ln \left[\frac{l_s^2}{R^2 \left(1 - \frac{T}{T_H}\right)} \right] . \tag{33.66}$$

Again we see that the tensor spectrum is approximately scale invariant, $n_T - 1 \simeq 0$, but now we find the tilt

$$n_T - 1 = \frac{d \ln P_h(k)}{d \ln k} = \left(-\frac{1}{1 - \frac{T}{T_H}} + 2 \right) k \frac{d}{dk} \frac{T(k)}{T_H} = -(n_s - 1) \frac{2 \frac{T(k)}{T_H} - 1}{2 - \frac{T(k)}{T_H}}$$
$$\simeq -(n_s - 1) > 0 \tag{33.67}$$

So the characteristic property of string gas cosmology is this blue tilt for the tensor spectrum, with the same deviation from scale invariance as the red tilt of the scalar spectrum. Inputting the experimentally observed scalar tilt from the Planck data, we expect that

$$n_T - 1 \simeq 0.03. \tag{33.68}$$

Moreover, the tensor to scalar ratio is now

$$r = \frac{P_h(k)}{P_\Phi(k)} \simeq \left(1 - \frac{T(k)}{T_H} \right)^2 \ln \left[\frac{1}{l_s^2 k^2} \left(1 - \frac{T(k)}{T_H} \right) \right] \ll 1 \,, \tag{33.69}$$

so a small ratio, consistent with observations, as well as being similar to string inflation models in that respect.

Important Concepts to Remember

- For a gas of Dp-branes, the wrapping modes have equation of state $w_p = -p/d$, generalizing the one for the strings.
- The addition of p-branes to strings means that we can have a hierarchy of large dimensions: besides the 3 very large ones, we can have two smaller but still large ones for 2-branes, etc.
- Two-point correlators of the energy-momentum tensor can be expressed in terms of thermodynamic quantities: for $\delta\rho$ in terms of C_V, and for spatial components in terms of $\partial P/\partial V|_T$.
- The thermodynamics of a gas of strings allows us to calculate the two-point correlators of energy-momentum tensor, and from them the power spectra of scalar (for $\delta\rho$) and tensor (for the spatial components) modes.
- Besides a small tensor-to-scalar ratio $r \ll 1$, which is also possible in inflation (especially appears in string inflation), we obtain the characteristic $n_T - 1 = -(n_s - 1)$.

Further reading: The spectrum of perturbations in string gas cosmology was found in [61], and the brane has generalization was defined in [62]. Good reviews for string gas and brane gas cosmology are [58, 59] and the recent one [60].

Exercises

(1) What order is the Hagedorn phase transition?
(2) What is the equation of state near the Hagedorn transition?
(3) Use experimental data to constrain the function $T(k)$.

(4) Show that for the metric

$$ds^2 = a^2(\eta)\left\{-(1+2\Phi(x,\eta))d\eta^2 + [(1-2\Phi(x,\eta))\delta_{ij} + h_{ij}]dx^i dx^j\right\}\ , \tag{33.70}$$

the Einstein equations give

$$\nabla^2\Phi = 4\pi G_N \delta\rho. \tag{33.71}$$

and

$$\nabla^2 h_{ij} = 4\pi G_N \delta T^i{}_j\ , \tag{33.72}$$

Chapter 34
Chameleon Scalars and String Theory

In this chapter we will study chameleon scalars and their embedding in string theory. It was usually thought the fifth force experiments conducted on Earth, or in the Solar System, would prevent the existence of light scalars.

However, the important observation of Khoury and Weltman, who developed the chameleon idea, was that we can have a mass for the scalar that depends on the local mass density. In a dense environment like the Earth, the chameleon will be very massive, so will elude the usual fifth force constraints. Moreover, we have a "thin-shell" effect, where for a very massive body only a thin shell surrounding it effectively interacts via the scalar force, which allows the chameleon to avoid the constraints on the Newtonian motion.

But on large cosmological scales, the scalar can be extremely light, so that it can have important consequences. As a plus, in string theory in principle we wouldn't need to "stabilize the moduli" (have large masses for them nonperturbatively), but rather they could be chameleons.

34.1 Chameleon (and Symmetron) Models

The action relevant for the chameleon models (as well as for the symmetron variant) is

$$S = \int d^4x\sqrt{-g}\left[\frac{M_{\rm Pl}^2}{2}R[g] - \frac{1}{2}g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi - V(\phi)\right] + \int d^4x\mathcal{L}_m[\tilde{g},\psi]\,, \tag{34.1}$$

where $\mathcal{L}_m$ is a matter action, depending on matter fields ψ, but which couples to the conformally-related metric

$$\tilde{g}_{\mu\nu} = A^2(\phi)g_{\mu\nu}. \tag{34.2}$$

Here $g_{\mu\nu}$ is the Einstein frame metric. The equation of motion for the scalar field (KG) is

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_34

$$\Box_g \phi - V_{,\phi} + A^3(\phi) A_{,\phi} \tilde{T} = 0 \,, \tag{34.3}$$

where

$$\tilde{T} = \tilde{T}_{\mu\nu} \tilde{g}^{\mu\nu} \tag{34.4}$$

is the trace of the matter (or Jordan) frame energy-momentum tensor

$$\tilde{T}_{\mu\nu} \equiv -\frac{2}{\sqrt{-\tilde{g}}} \frac{\delta \mathcal{L}_m}{\delta \tilde{g}^{\mu\nu}}. \tag{34.5}$$

This tensor is conserved in the $\tilde{g}_{\mu\nu}$ (Jordan) frame as usual, i.e.,

$$\tilde{\nabla}_\mu \tilde{T^\mu}_\nu = 0. \tag{34.6}$$

We define the density in the Jordan frame as usual, by

$$\tilde{T^\mu}_\nu = \text{diag}(-\tilde{\rho}, \tilde{P}, \tilde{P}, \tilde{P}) \,, \tag{34.7}$$

and for dust matter ($P \simeq 0$) we have

$$\tilde{T^\mu}_\mu \simeq -\tilde{\rho}. \tag{34.8}$$

The energy-momentum tensor in the Jordan frame can be related to the one in Einstein frame by

$$\begin{aligned} \tilde{T}_{\mu\nu} &= \frac{1}{A^4} \frac{-2}{\sqrt{-g}} A^2 \frac{\delta \mathcal{L}_m}{\delta g^{\mu\nu}} \Rightarrow \\ \tilde{T^\mu}_\nu &= \frac{1}{A^4} \left(-\frac{2}{\sqrt{-g}} g^{\mu\rho} \frac{\delta \mathcal{L}_m}{\delta g^{\rho\nu}} \right) \Rightarrow \\ -\tilde{\rho} &= \frac{1}{A^4} T^\mu{}_\mu. \end{aligned} \tag{34.9}$$

On the other hand, the zero component of the conservation equation becomes

$$\begin{aligned} \tilde{\nabla}_\mu \tilde{T^\mu}_0 &= \tilde{\nabla}_0 \tilde{T^0}_0 + \tilde{\nabla}_i \tilde{T^i}_0 \\ &= \nabla_0 \tilde{T^0}_0 + \tilde{\Gamma}^i_{i0} \tilde{T^i}_0 = \nabla_0 \tilde{T^0}_0 + \left(3\frac{\dot{A}}{A} + \Gamma^i_{i0} \right) \tilde{T^0}_0 \\ &= \nabla_0 \tilde{T^0}_0 + \left(3\frac{\dot{A}}{A} + \nabla_i \right) \tilde{T^i}_0. \end{aligned} \tag{34.10}$$

Then the conservation equation becomes

$$\left(\nabla_0 + 3\frac{\dot{A}}{A}\right)\tilde{\rho} = 0. \tag{34.11}$$

That means that if we define the density ρ by

$$\rho \equiv A^3\tilde{\rho}\ , \tag{34.12}$$

then it will be conserved (time independent in a non-expanding Universe) in Einstein frame. In particular, it will be independent of ϕ. Moreover we obtain

$$T^{\mu}{}_{\mu} = A\rho\ , \tag{34.13}$$

which means that the equation of motion for ϕ is

$$\Box_g\phi = V_{,\phi} + A_{,\phi}\rho. \tag{34.14}$$

Finally, that means that we have an *effective potential*

$$V_{\rm eff}(\phi) = V(\phi) + \rho A(\phi). \tag{34.15}$$

For a spherically symmetric solution, the equation of motion for the chameleon is

$$\frac{d^2\phi}{dr^2} + \frac{2}{r}\frac{d\phi}{dr} = V_{,\phi} + A_{,\phi}\rho. \tag{34.16}$$

We have now the interesting possibility that even though $V(\phi)$ might not have a minimum, $V_{\rm eff}(\phi)$ can, and the effective mass at the minimum of the effective potential is

$$m^2_{\rm eff} = V_{,\phi\phi}(\phi_{\rm min}) + A_{,\phi\phi}(\phi_{\rm min})\rho. \tag{34.17}$$

Since the ρ term increases generically with ϕ, we need $V_{,\phi} < 0$ over the period of interest in order to have a minimum for the effective potential (the sum of the two terms). For stability, i.e., in order to have a positive mass squared, we need $V_{,\phi\phi} > 0$. Moreover, for increasing of the effective mass with ρ (so that the mass on Earth is bigger than outside it), we need $V_{,\phi\phi\phi} < 0$.

Examples of potentials that obey these constraints are the inverse power law,

$$V \sim \frac{1}{\phi^n}\ ,\quad n > 0\ , \tag{34.18}$$

the exponential

$$V \sim e^{-a\phi}\ , \tag{34.19}$$

or more complicated exponentials.

A typical coupling function $A(\phi)$, of the type that also is natural in string theory, is

$$A(\phi) \simeq e^{g\frac{\phi}{M_{\rm Pl}}} \simeq 1 + g\frac{\phi}{M_{\rm Pl}}. \tag{34.20}$$

The main difference (with respect to generic chameleons) in the case of the symmetron model is an $A(\phi)$ that doesn't have a linear term, only a quadratic one, i.e., $A(\phi) \simeq (\phi - \phi_0)^2$.

34.2 Thin-Shell Effect

Consider a spherically symmetric object of radius $\mathcal{R}$, density $\rho_{\rm in}$, associated with a field $\phi_{\rm min-in}$, embedded in a medium with density $\rho_{\rm out}$, associated with a field $\phi_{\rm min-out}$.

Then for a sufficiently large object, ϕ will start in the core at $\phi_{\rm min-in}$, and it will start to deviate from this value at $\mathcal{R}_{\rm roll}$. The deviation can be approximated to be quadratic, so we find

$$\begin{aligned}\phi &\simeq \phi_{\rm min-in}; \quad 0 < r < \mathcal{R}_{\rm roll}\\ &\simeq \phi_{\rm min-in} + \frac{g}{2M_{\rm Pl}}\rho_{\rm min}(r - \mathcal{R}_{\rm roll})^2; \quad \mathcal{R}_{\rm roll} < r < \mathcal{R}.\end{aligned} \tag{34.21}$$

By matching with the outside value for the scalar, we find

$$\frac{\Delta\mathcal{R}}{\mathcal{R}} = \frac{\mathcal{R} - \mathcal{R}_{\rm roll}}{\mathcal{R}} = \frac{\phi_{\rm min-out} - \phi_{\rm min-in}}{6gM_{\rm Pl}\Phi}. \tag{34.22}$$

Here

$$\Phi \equiv \frac{\rho_{\rm min}\mathcal{R}^2}{6M_{\rm Pl}^2} \tag{34.23}$$

is the Newton potential at the surface (compared to infinity, outside).

Finally then, the scalar field outside the sphere is a *screened* one, by a factor of $\Delta\mathcal{R}/\mathcal{R}$,

$$\phi_{\rm screened} \simeq -\frac{g}{4\pi M_{\rm Pl}}\frac{\Delta\mathcal{R}}{\mathcal{R}}\frac{Me^{-m_{\rm min-out}r}}{r} + \phi_{\rm min-out}. \tag{34.24}$$

We leave this as an exercise to prove.

In turn, that means that the effective coupling between two large bodies is

$$g_{\rm eff}^2 = g^2\left(\frac{\Delta\mathcal{R}}{\mathcal{R}}\right)_1\left(\frac{\Delta\mathcal{R}}{\mathcal{R}}\right)_2. \tag{34.25}$$

Concluding this subsection and experimental constraints, generically we have

$$m_{\text{eff}}^2(\text{Earth}) \gg V''(\phi_{\text{out}}) \ , \tag{34.26}$$

which means that the chameleon is quite massive on Earth, and avoids fifth force detection.

Moreover, in the Solar System, we have

$$\left(\frac{\Delta\mathcal{R}}{\mathcal{R}}\right)_{\text{planet}} \ll 1 \ , \tag{34.27}$$

which means that we don't see the scalar force in the Solar System because of the thin-shell effect: the effective coupling is very small.

On the other hand, the chameleon is left to be light on cosmological scales, which means that we can have cosmological effects.

34.3 String Theory Embedding

The chameleon idea is very welcome in string theory. Indeed, we saw that we have many moduli that need stabilization, and when we try to do stabilize all, we usually spoil inflation. It would be of great help therefore if some of these moduli would not need to be stabilized, but rather would be chameleons. Then they could be hidden on Earth and in the Solar System, but would still be detectable cosmologically.

One way to embed the chameleons in string, or rather string-inspired, models, takes ϕ as the volume modulus (related to the volume of the compact dimensions) in a KK dimensional reduction.

Under KK dimensional reduction, we generically have

$$ds_D^2 = R^2 ds_d^2 + g_{mn} dx^m dx^n + \cdots \ , \tag{34.28}$$

where

$$ds_d^2 = g_{\mu\nu} dx^\mu dx^\nu \tag{34.29}$$

is the d-dimensional Einstein frame.

Here

$$R = \Delta^{-\frac{1}{d-2}}; \quad \Delta = \sqrt{\det g_{mn}}. \tag{34.30}$$

Under the important assumption that the matter action is simple (becomes the usual matter) in the ds_D^2 Jordan frame rather than the Einstein frame (which fact is not obvious a priori in KK compactifications), we have the identification

$$A^2(\phi) = R^2 = \Delta^{-\frac{2}{d-2}} = (\det g_{mn})^{-\frac{2}{d-2}}. \tag{34.31}$$

Choosing the physical case of $D = 10$, $d = 4$, we have

$$A^2(\phi) = R^2 = \Delta^{-1} = \frac{1}{V_6 M_{\rm Pl}^6} = \frac{1}{r^6 M_{\rm Pl}^6}. \tag{34.32}$$

In general, in KK compactifications in string theory, the canonical scalar is related to R by

$$\phi = \frac{M_{\rm Pl}}{g} \ln \frac{R}{R_*} , \tag{34.33}$$

for some dimensionless coupling g and some constant R_*.

But for the volume modulus there must be some nonperturbative potential stabilizing it, translating into a potential for R.

Around the minimum (stabilization point) of this potential, we expand:

- for $R \gtrsim R_*$,
$$V(R) = M_{\rm Pl}^4[-\alpha(R - R_*) + \beta(R - R_*)^2]. \tag{34.34}$$

The minimum of the potential is at

$$R_{\rm min} = R_* + \frac{\alpha}{2\beta}. \tag{34.35}$$

- for $R \lesssim R_*$, the potential is a steep exponential,

$$V(R) = M_{\rm Pl}^4 v[e^{\gamma(R^{-k} - R_*^{-k})} - 1] , \tag{34.36}$$

where $\alpha, \beta, \gamma, k, v > 0$ are dimensionless. By matching the two behaviour at $R = R_*$, we find the relation

$$\alpha = v \frac{k\gamma}{R_*^{k-1}}. \tag{34.37}$$

Example: KKLT

As an example of this postulated behaviour, we show that the KKLT model is of this type.

Consider a complex volume modulus ρ, then

$$\sigma = {\rm Im}\rho = \frac{M_s^4 r^4}{g_s} = \frac{M_{10}^4 r^4}{2\sqrt{\phi} g_s} , \tag{34.38}$$

where M_{10} is the 10 dimensional Planck scale and $M_s = 1/(2\pi\sqrt{\alpha'})$ is the string scale.

But if instead, we would have the relation

$$\sigma = \frac{M_{\rm Pl}^4 r^4}{g_s} = \frac{R^{-4/3}}{g_s} \,, \tag{34.39}$$

which differs from the above by a large factor a that can be absorbed in the exponent of the KKLT superpotential (though now a doesn't appear naturally anymore), then we obtain the picture advocated above.

As we already saw, the Kähler potential for the volume modulus is

$$K = -3\ln[-i(\rho - \bar{\rho})] \,, \tag{34.40}$$

and the KKLT superpotential is

$$W = W_0 + Ae^{ia\rho}. \tag{34.41}$$

Moreover, we can obtain also $a < 0$, which is what we will need, by modifying the standard KKLT scenario, as we saw.

We can prove (we leave this as an exercise) that the supersymmetric KKLT potential near its minimum can be approximated by

$$V_{\rm KKLT} - V_0 \simeq \frac{A^2 a^2}{6M_{\rm Pl}^2 \sigma_*} \left[e^{2|a|\sigma} - e^{2|a|\sigma_0}\right]. \tag{34.42}$$

34.4 Experimental Constraints

We can now put the experimental constraints on this model. Fifth force experiments on Earth give

$$g_{\rm eff}^2 \lesssim 10^{-3} \,, \tag{34.43}$$

which means

$$\frac{3\Delta\mathcal{R}}{\mathcal{R}} = \frac{\phi_{\rm vac} - \phi_{\rm test-mass}}{2gM_{\rm Pl}\Phi_{\rm test-mass}} \lesssim 10^{-3/2} \,, \tag{34.44}$$

that finally implies

$$\phi_{\rm vac} - \phi_{\rm test} \lesssim 10^{-29} M_{\rm Pl}. \tag{34.45}$$

After some algebra, this is shown to imply

$$R_*^k \lesssim 10^{-30}\gamma k. \tag{34.46}$$

Taking the values close to the constraint, we obtain the range of the chameleon in various environments as

$$m_{\rm eff}^{-1} \lesssim \frac{0.2\,{\rm mm}}{\sqrt{\rho[{\rm g/cm^3}]}} \,, \tag{34.47}$$

which means that in the atmosphere

$$m_{\rm atm}^{-1} \lesssim 1\,{\rm cm}\ , \tag{34.48}$$

and in the Solar System

$$m_{\rm Solar\ system}^{-1} \lesssim 10^5\,{\rm Km}. \tag{34.49}$$

But moreover, in the Solar System, the thin-shell factors are extremely small. Even for the Moon we find

$$\left(\frac{3\Delta\mathcal{R}}{\mathcal{R}}\right)_{\rm Moon} \lesssim 10^{-10}. \tag{34.50}$$

We need to impose that the galaxy itself is screened, otherwise its local field will not be given by its density. Then

$$\left(\frac{3\Delta\mathcal{R}}{\mathcal{R}}\right)_{\rm galaxy} < 1\ , \tag{34.51}$$

which in the end imposes the constraint

$$\frac{\alpha}{\beta} \lesssim 10^{-6} R_*. \tag{34.52}$$

On the other hand, the cosmological density must be low, so that ϕ doesn't lie on the steep exponential part. This, after some algebra, implies

$$\alpha \geq 10^{-120} R_*^{-1}. \tag{34.53}$$

Finally, we find that the mass of the chameleon on cosmological scales obeys

$$m_{\rm cosmo} = \sqrt{2} g M_{\rm Pl} \sqrt{\beta} R_* \gtrsim 10^3 H_0. \tag{34.54}$$

Applying this for KKLT, we find

$$|V_0| \simeq 10^{-174} M_{\rm Pl}^4; \quad W_0 \simeq 10^{-42} M_{\rm Pl}^3; \quad A \sim M_{\rm Pl}^3 e^{-10^{30}}. \tag{34.55}$$

However, the last constraint doesn't look so absurd when it is rewritten as

$$A e^{-a\sigma} = M_{\rm Pl}^3 e^{|a|(\sigma-\sigma_0)}\ , \tag{34.56}$$

with $\sigma_0 \sim 10^{30}$.

However, it turns out that we need to apply the constraints directly to KKLT, the general model not being a good enough approximation. Then we find instead

$$
\begin{aligned}
&|a|\sigma_{\min} > 10^6 \\
&|V_0| \geq \frac{10^{-120}}{|a|\sigma_{\min}} M_{\rm Pl}^4 \gtrsim 10^{-159} M_{\rm Pl}^4 \\
&W_0 \gtrsim 10^{-30} M_{\rm Pl}^3 .
\end{aligned}
\tag{34.57}
$$

Those are not that important modifications. However, we also find that now

$$
m_{\rm cosmo} \gtrsim 10^{15} H_0 \ , \tag{34.58}
$$

which is an important difference, since now the chameleon is not relevant for cosmological distances anymore. But in fact it turns out that with two KKLT exponentials in the superpotential instead of one (very close in value to each other), we can achieve the general formula of $m_{\rm cosmo} \geq 10^3 H_0$.

Important Concepts to Remember

- Chameleon scalars are scalars with mass that depends on the local density, that are massive on Earth and avoid detection.
- They also have a thin-shell effect: only a thin shell of a massive body interacts via the scalar, avoiding detection for planets in the Solar System.
- The chameleon is obtained when the matter couples to $\tilde{g}_{\mu\nu} = A^2(\phi) g_{\mu\nu}$, resulting in an effective potential $V_{\rm eff} = V + \rho A(\phi)$.
- The density $\rho = A^3 \tilde{\rho}$ is conserved in Einstein frame.
- The thin-shell effect gives a screening coefficient of $3\Delta\mathcal{R}/\mathcal{R} = \Delta\phi/(6gM_{\rm Pl}\Phi)$, with Φ the Newton potential at the surface.
- We can embed in string theory with ϕ being the volume modulus.
- To obtain a chameleon, consider a potential for stabilizing R that has a minimum, and a steep exponential on one side, $e^{\gamma R^{-k}}$.
- The KKLT model gives such an example, with $k = 4/3$.
- Generically, we obtain $m_{\rm cosmo} \gtrsim 10^3 H_0$. In simple KKLT, we obtain instead $m_{\rm cosmo} \gtrsim 10^{15} H_0$, but with two exponentials we can get the general formula.

Further reading: The chameleon scalars were defined in [63, 64]. The embedding in string theory described here was done in [65].

Exercises

(1) Check that the screened solution outside the spherical mass source is

$$
\phi_{\rm screened} \simeq -\frac{g}{4\pi M_{\rm Pl}} \frac{\Delta\mathcal{R}}{\mathcal{R}} \frac{M e^{-m_{\rm min-out} r}}{r} + \phi_{\rm min-out} . \tag{34.59}
$$

(2) Prove that we can approximate the KKLT potential around the minimum by

$$
V_{\rm KKLT} - V_0 \simeq \frac{A^2 a^2}{6 M_{\rm Pl}^2 \sigma_*} \left[e^{2|a|\sigma} - e^{2|a|\sigma_0} \right] . \tag{34.60}
$$

(3) Check and explain the origin of the difference in chameleon mass on cosmological scales for the KKLT model and for the generic model with a minimum and a steep exponential on one side.
(4) Is the thin-shell effect, encapsulated by the solution at exercise 1, still valid for the symmetron case, with $A(\phi) = (\phi - \phi_0)^2$?

Chapter 35
Axion Inflation and Axion Monodromy from String Theory

In this chapter, we will learn about axion inflation in field theory, how axions appear in string theory, and a specific set-up for string inflation that has been very popular recently, axion monodromy.

35.1 Axions in Field Theory

In a nonabelian gauge theory, we can have instantons, self-dual solutions of the Yang–Mills equations in Euclidean space (satisfying $F^a_{\mu\nu} = \pm * F^a_{\mu\nu} \equiv \frac{1}{2}\epsilon_{\mu\nu\rho\sigma}F^{a\,\rho\sigma}$), carrying the topological number called the second Chern number (or Pontryagin number, or instanton number)

$$c_2 = \int_{M_4} \frac{1}{32\pi^2}\,\mathrm{Tr}[F_{\mu\nu} * F^{\mu\nu}] \in \mathbb{Z}. \tag{35.1}$$

The instanton action is $S = 8\pi^2/g^2$, and at the quantum level it gives transitions between states with different *winding numbers* n for the gauge configuration on the spatial $\mathbb{R}^3$. Therefore gauge fields with a given winding number n are in general not a good vacuum of the theory by itself. A good vacuum is a superposition of the states $|n\rangle$ with winding number n, parametrized by a constant θ as

$$|\theta\rangle = \sum_{n\geq 0} e^{-in\theta}|n\rangle. \tag{35.2}$$

This is called the θ vacuum, and here θ is an a priori free parameter. The effects of this parameter can be repackaged as a parameter multiplying the topological term giving the instanton in the Lagrangian,

$$\mathcal{L}_{\mathrm{top}} = \frac{\tilde{\theta}g^2}{32\pi^2}F^a_{\mu\nu} * F^{a\,\mu\nu}. \tag{35.3}$$

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_35

Actually, the parameter in the Lagrangian contains, besides θ, a quantum contribution from fermions coupled to it,

$$\tilde{\theta} = \theta + arg(\det M) \ , \tag{35.4}$$

where M is the fermion mass matrix.

But in the real world, in QCD, it turns out that θ is measured (bound) to be extremely small, less than 10^{-10}. To explain why this parameter is so small, the solution proposed by Peccei and Quinn is to first promore θ to a pseudoscalar field $\theta(x)$ and then assume there is a global chiral symmetry (called Peccei-Quinn symmetry) that acts on it as a shift symmetry, $\theta \to \theta$+const., and on the fermions, for instance by multiplying the mass term by a γ_5 phase, effectively $M \to e^{i\delta\theta\gamma_5} M$, and which is spontaneously broken at some scale f_{PQ}. The Nambu-Goldstone boson for its spontaneous breaking is the pseudoscalar field called the axion $\theta(x) = a(x)$, and the breaking is due to instanton effects.

This chiral symmetry is anomalous since we have fermions charged under it, and in the presence of an instanton it generates the topological term. Indeed, the anomaly of a global symmetry in a gauge theory is

$$\partial_\mu j_\mu = \frac{C_a g^2}{16\pi^2} \mathrm{Tr}[F_{\mu\nu} * F^{\mu\nu}] \ , \tag{35.5}$$

where C_a depends on the number and charges of the fermions. Therefore in the presence of an instanton it generates a contribution to the action given by the above, times the axion $\theta(x) = a(x)$.

This in turn means a potential term for the axion, fixing it to have the total VEV equal to zero, the potential coming from the instanton contributions, and breaking the symmetry at some scale, in particular giving a mass Λ^2_{QCD}/f_{PQ}. Here Λ_{QCD} is the dynamically generated scale of the gauge theory, and $f_{PQ} = f$ is also called *axion decay constant*.

The Lagrangian for the axion can thus be taken to be

$$\mathcal{L}(a) = -\frac{1}{2} f^2 (\partial_\mu a)^2 - \Lambda^4 [1 - \cos(a)] + \cdots \ , \tag{35.6}$$

where the dots represent higher derivative corrections and multi-instanton contributions, or, by defining a canonically normalized field $\phi = fa$,

$$\mathcal{L}(\phi) = -\frac{1}{2}(\partial_\mu \phi)^2 - \Lambda^4 \left[1 - \cos\frac{\phi}{f}\right] + \cdots \simeq -\frac{1}{2}(\partial_\mu \phi)^2 - \frac{\Lambda^4}{2f^2}\phi^2 + \cdots \tag{35.7}$$

Then the axion has now just a periodicitiy $a \to a + 2\pi$, or $\phi \to \phi + 2\pi f$, instead of the general shift symmetry.

35.2 Axions in String Theory

In string theory, we look for fields $a(x)$ that have shift symmetries $a \to a+\text{const.}$, and which then can be broken by nonperturbative effects to constant shifts.

There are many such fields in string compactifications, arising from the integration of p-form potentials over p-cycles in the compact space.

Specifically:

- in type IIA string theory we would have the NS-NS 2-form $B_{\mu\nu}$ integrated over 2-cycles Σ_2^I, the RR one-form C_μ integrated over one-cycles Σ_1^I, and the RR 3-form $C_{\mu\nu\rho}$ integrated over 3-cycles Σ_3^I.
- in type IIB, where we will concentrate, we have again the NS-NS 2-form $B_{\mu\nu}$, but now also the R-R 2-form $C_{\mu\nu}$ integrated over 2-cycles Σ_2^I, the R-R 4-form $C_{\mu\nu\rho\sigma}$ integrated over 4-cycles Σ_4^I. We also have 3 universal axions, 4 dimensional pseudoscalars (not integrated over cycles), the IIB axion C, as well as the *4-dimensional* duals b and c of $B_{\mu\nu}$ and $C_{\mu\nu}$ ($dB_2 = H_3 = *_4 H_1$, $H_1 = db$, $dC_2 = F_3 = *_4 F_1$, $F_1 = dc$).

We then have, in type IIB,

$$b_I = \frac{1}{2\pi\alpha'}\int_{\Sigma_2^I} B_2 \ , \quad c_I = \frac{1}{2\pi\alpha'}\int_{\Sigma_2^I} C_2 \ , \quad \tilde{c}_I = \frac{1}{2\pi\alpha'}\int_{\Sigma_4^I} C_4 \ , \tag{35.8}$$

together with C, b and c. Together, these are called a (the dimensionless axion defined before, with period 2π).

Generically, the p-forms have gauge invariances, $\delta C_p = d\Lambda_{p-1}$, which leads, under integration on a cycle, to $\delta\int_{\Sigma_p} C_p = \int_{\Sigma_p} d\Lambda_{p-1}$, which can be taken to be an arbitrary constant, under a suitable choice of gauge parameter $(p-1)$-form Λ_{p-1}. For instance, for $C_1 = d\lambda$, we have

$$\delta c = \delta \oint_{S^1} d\lambda = \int_{S^1} dx^\mu \partial_\mu \lambda = \lambda(x + 2\pi R) - \lambda(x) \ , \tag{35.9}$$

which can be chosen to be what we want.

Thus classically, we have $\delta a = const.$ as a symmetry. But at the quantum level, we can have *worldsheet, or D-brane, instantons*. A D-p-brane is a source for a C_{p+1}, so if an *Euclidean* (worldsheet signature) D-brane wraps that same cycle, there is a priori a contribution to the axion action, that breaks the shift symmetry. The same thing happens for a fundamental string Euclidean worldsheet wrapping a 2-cycle. Because of the Euclidean signature, these are both instantonic solutions.

Let us describe this last example in a bit more detail. The action for the string worldsheet contains the term

$$S_B = -\frac{1}{4\pi\alpha'}\int_{\Sigma_2^I} d^2\sigma \epsilon^{ab}\partial_a X^M \partial_b X^N B_{MN}(X) = -\frac{1}{2\pi\alpha'}\int_{\Sigma_2^I} B_2 \equiv -b(x^\mu)^I \ , \tag{35.10}$$

where $b(x^\mu)$ is the dimensionless axion ($a(x^\mu)$ in the general notation), and we have denoted $M = (\mu, i)$, with μ being 4 dimensional (noncompact) index and i compact index. This term contributes to the axion action. We can expand the coordinates X^M around some point $X^M_{(0)}$.

Considering only $X^M_{(0)}$, so that the B-field is a constant, both in the worldsheet and in spacetime, we get that the contribution is

$$\delta S_B = -\frac{1}{4\pi\alpha'} \int_{\Sigma_2^I} \partial^2\sigma \partial_a \left(\epsilon^{ab} X^M \partial_b X^N B_{MN}(X^K_{(0)})\right) , \tag{35.11}$$

and this vanishes unless there is a boundary or a topologically nontrivial cycle for Σ_2^I. The terms involving the fluctuations in X^M will come with derivatives, $\partial_P B_{MN}(X_{(0)})$, translating in $\partial_\mu b$ terms in spacetime, which respect the shift symmetry.

But the worldsheet instantons, worldsheets wrapping nontrivial cycles Σ_2^I, give instantonic contributions to the action. B_2 couples to the Kähler form J for the 2-cycle into a complex field, as we saw before, so $(J + iB_2)$ is integrated over Σ_2, for an instanton action

$$S_{\text{instanton}} = \exp\left(-\frac{1}{2\pi\alpha'} \int_{\Sigma_2^I} (J + iB_2)\right) = e^{-ib^I}. \tag{35.12}$$

This is a contribution to the superpotential for the b axion, which has therefore a periodicity $b \to b + 2\pi$, as expected.

To determine the axion decay constant f, we must calculate the kinetic term for the axion. In the case of b, the 10 dimensional kinetic term, coming from the type II supergravity action with coefficient $-1/(4\kappa^2)$, is[1]

$$S_{\text{B,sugra}} = -\frac{1}{2(2\pi)^7 g_s^2 \alpha'^4} \int_{M_{10}} d^{10}x \sqrt{-g} |dB_2|^2. \tag{35.13}$$

We expand B_2 in a basis Ω^I of 2-forms dual to the cycles Σ_2^I,

$$B_2 = \sum_I b_I(x)\Omega^I , \tag{35.14}$$

where $b_I(x)$ are the 4-dimensional fields, and the 2-forms Ω^I are normalized by (note that B_2 is dimensionless, and b_I are also, so from this, and the instanton action below, the normalization must be)

[1] If we call, as usual, the coefficient of the Einstein action as $1/(2\kappa^2)$, then one can find that the D-string tension is $T_{D1} = 4\pi^{5/2}\alpha'/\kappa$, from matching the D-brane action against a string theory calculation. The string tension is $T_{F1} = 1/(2\pi\alpha')$, so their ratio is $T_{F1}/T_{D1} = \kappa/(8\pi^{7/2}\alpha'^2)$, and is usually defined to be equal to g_s, leading to $2\kappa^2 = (2\pi)^7 g_s^2 \alpha'^4$.

$$\int_{\Sigma_2^I} \Omega^J = 2\pi\alpha' \delta_I^J . \tag{35.15}$$

Then, substituting in the supergravity action (note that $|F_p|^2 \equiv \frac{1}{p!} g^{\mu_1\nu_1} \dots g^{\mu_p\nu_p} F_{\mu_1\dots\mu_p} F_{\nu_1\dots\nu_p}$), we obtain

$$S_{\text{B,sugra}} = -\frac{1}{2} \int d^4x \sqrt{-g} g^{IJ} (\partial_\mu b_I)(\partial^\mu b_J) \ , \tag{35.16}$$

where the sigma model metric on the axion space is (from the above, with $1/3! = 1/6$ coefficient)

$$g^{IJ} \equiv \frac{1}{6(2\pi)^7 g_s^2 \alpha'^4} \int_{K_6} \Omega^I \wedge *_6 \Omega^J . \tag{35.17}$$

But, from the Einstein action, which in string theory is found to be ($2\kappa^2$ was considered before)

$$S = \frac{1}{2\kappa^2} \int_{M_{10}} R = \frac{1}{(2\pi)^7 g_s^2 \alpha'^4} \int_{M_{10}} R = \frac{M_{\text{Pl}}^2}{2} \int_{M_4} R_{(4)} + \cdots \ , \tag{35.18}$$

we get that (M_{Pl} is the *4-dimensional* Planck scale)

$$M_{\text{Pl}}^2 = \frac{2V}{(2\pi)^7 g_s^2 \alpha'^4} \ , \tag{35.19}$$

where V is the volume of the compact space K_6. For an isotropic compactification with scale L, so that $V = L^6$, and because of our normalization of Ω^I ($\int_{\Sigma_2^I} \Omega^J = 2\pi\alpha' \delta_I^J$), we find $\int_{K_6} \Omega^I \wedge *_6 \Omega^J \simeq L^2 \alpha'^2 \delta^{IJ}$. Finally, diagonalizing the metric g^{IJ} to $f^2 \delta^{IJ}$, we obtain

$$g^{IJ} \simeq \frac{M_{\text{Pl}}^2 \alpha'^2}{12 L^4} \delta^{IJ} \Rightarrow \frac{f^2}{M_{\text{Pl}}^2} \simeq \frac{\alpha'^2}{12 L^4} . \tag{35.20}$$

Since in order to trust computations we need $L \gg \sqrt{\alpha'}$, it means we have $f \ll M_{\text{Pl}}$.

Axionic Inflation from String Theory

But for n_s to be compatible with "natural inflation" coming from the axion action, the Planck experiment (see [76], Fig. 11) finds that one needs $f \gtrsim 10 M_{\text{Pl}}$. Note that the result depends very drastically on priors; the cited formula is for a uniform prior for $\log f$. But as we saw, super-Planckian values for f are problematic.

So it seems that the simplest models with the string theory axions don't work, and we need to do something else. There are 3 main solutions to this: a fine tuning in the case of two axions similar to "racetrack" inflation fine tuning, a version of N-flation

(inflation with a radial field in an N-dimensional space of inflatons), and the best candidate from string theory, axion monodromy inflation.

The first two cases amount to having f's that are sub-Planckian for each axion, but the effective axion has a trans-Planckian decay constant, and therefore field range.

35.3 "Racetrack" with Two Axions

It is a known fact that if we fine-tune the sum of two exponentials or trigonometric functions (which are exponentials of imaginary argument), we can obtain a function that has a longer period, so it looks "flatter". This is basically the mechanism for "racetrack" inflation, which is an inflationary potential made up of two close by exponentials.

A similar mechanism can be constructed for axion potentials, where the some of two cosine axionic potentials will lead to a potential with a much longer period, i.e., range or axion decay constant f.

One considers two axionic scalars, each of which couples to a linear combination of two confining nonabelian gauge groups, that lead to a cosine potential. That is, the coupling term in the action is (from the form (35.3) for a rescaled dimensionless axion $\tilde{\theta} = \varphi/f$, where φ has a canonical kinetic term, and for two different group factors a, b)

$$S_{\text{coupling}} = \int d^4x \sum_{i=1,2} \frac{\varphi_i}{f_i} \left(\frac{c_{ia}}{32\pi^2} \operatorname{Tr}\left[F^{(a)} \wedge F^{(a)}\right] + \frac{c_{ib}}{32\pi^2} \operatorname{Tr}\left[F^{(b)} \wedge F^{(b)}\right] \right) , \tag{35.21}$$

and it leads at low energies to the axion potential

$$V = \Lambda_a^4 \left[1 - \cos\left(c_{1a} \frac{\varphi_1}{f_1} + c_{2a} \frac{\varphi_2}{f_2} \right) \right] + \Lambda_b^4 \left[1 - \cos\left(c_{1b} \frac{\varphi_1}{f_1} + c_{2b} \frac{\varphi_2}{f_2} \right) \right] . \tag{35.22}$$

Consider now that we have approximately (but to a high precision)

$$C \equiv \frac{c_{2a}}{c_{1a}} \simeq \frac{c_{2b}}{c_{1b}} . \tag{35.23}$$

Then the potential becomes

$$V \simeq \Lambda_a^4 + \Lambda_b^4 - \Lambda_a^4 \cos \frac{c_{1a}}{f_1} \left(\varphi_1 + C \frac{f_1}{f_2} \varphi_2 \right) - \Lambda_b^4 \cos \frac{c_{1b}}{f_1} \left(\varphi_1 + C \frac{f_1}{f_2} \varphi_2 \right) . \tag{35.24}$$

As we can see, the potential is now a function of only a linear combination of the fields, $\varphi_1 + C \frac{f_1}{f_2} \varphi_2$, and is independent of the transverse linear combination, which is canonically normalized as

$$\xi \equiv \frac{c_{1a}\varphi_2 f_2 - c_{2a}\varphi_1 f_1}{\sqrt{c_{2a}^2 f_1^2 + c_{1a}^2 f_2}} \propto \varphi_2 - C\varphi_1 \frac{f_1}{f_2}. \tag{35.25}$$

Note that we have

$$(\partial\varphi_1)^2 + (\partial\varphi_2)^2 = \left(\frac{\varphi_1 + C\frac{f_2}{f_1}\varphi_2}{\sqrt{1 + C^2\frac{f_1^2}{f_2^2}}}\right)^2 + \left(\frac{\varphi_2 - C\frac{f_2}{f_1}\varphi_1}{\sqrt{1 + C^2\frac{f_1^2}{f_2^2}}}\right)^2, \tag{35.26}$$

and the second term is $(\partial\xi)^2$, whereas the first is the canonically normalized $\varphi_1 + C\frac{f_1}{f_2}\varphi_2$.

That means that ξ has no potential, therefore infinite range (infinite periodicity, or axion decay constant f). If we remember that the ratios of constants c_{ia} are not exactly equal, we can rewrite the potential in terms of the two modes defined above (we define now $C \equiv c_{2a}/c_{1s}$), as

$$\begin{aligned} V &= \Lambda_a^4\left[1 - \cos\left(\frac{c_{1a}\sqrt{1 + C^2\frac{f_1^2}{f_2^2}}}{f_1}\,\frac{\varphi_1 + C\frac{f_2}{f_1}\varphi_2}{\sqrt{1 + C^2\frac{f_1^2}{f_2^2}}}\right)\right] + \Lambda_b^4[1 - \\ &\quad - \cos\left(\frac{c_{1a}}{f_1}\,\frac{c_{1a}c_{1b}f_2^2 + c_{2a}c_{2b}f_1^2}{c_{1a}^2 f_2^2 + c_{2a}^2 f_1^2}\left(\varphi_1 + \frac{c_{2a}}{c_{1a}}\frac{\varphi}{f_2}\right) + \frac{(c_{1a}f_2\varphi_2 - c_{2a}f_1\varphi_1)(c_{2b}c_{1a} - c_{2a}c_{1b})}{c_{2a}^2 f_1^2 + c_{1a}^2 f_2^2}\right)\right] \\ &\equiv \Lambda_a^4\left[1 - \cos\left(\frac{c_{1a}\sqrt{1 + C^2\frac{f_1^2}{f_2^2}}}{f_1}\,\frac{\varphi_1 + C\frac{f_2}{f_1}\varphi_2}{\sqrt{1 + C^2\frac{f_1^2}{f_2^2}}}\right)\right] \\ &\quad + \Lambda_b^4\left[1 - \cos\left(\alpha\frac{\varphi_1 + C\frac{f_2}{f_1}\varphi_2}{\sqrt{1 + C^2\frac{f_1^2}{f_2^2}}} + \beta c_{1b}\frac{\xi}{f_\xi}\right)\right], \end{aligned} \tag{35.27}$$

where $\beta = 1$. We can find the above relation by putting arbitrary coefficients α, β and fixing them in order to get the same thing on both sides. Then, the axion decay constant for the mode ξ becomes

$$f_\xi = \frac{c_{1b}\sqrt{c_{2a}^2 f_1^2 + c_{1a}^2 f_2^2}}{|c_{2b}c_{1a} - c_{2a}c_{1b}|}, \tag{35.28}$$

so it can easily be made to be transplanckian, in particular $f_\xi > 10M_{\rm Pl}$.

35.4 N-Flation

Another known mechanism to extend the range of the inflaton so that it can become transplanckian is to consider N copies of the same scalar field, which leads to an effective field that is the radius in N dimensional field space, effectively

$$\Phi^2 \equiv \sum_{i=1}^{N} (\phi_i)^2 = N|\phi|^2 \Rightarrow \Phi = \sqrt{N}|\phi|. \tag{35.29}$$

This enhancing of the range by $\sqrt{N}$ due to the presence of N fields is called N-flation.

To be more precise, for an action

$$S = \sum_{i=1}^{N} \int d^4x \left[-\frac{1}{2}(\partial_\mu \phi_i)^2 - V_i(\phi_i) \right] , \tag{35.30}$$

calculating the KG equations, we see that each scalar field experiences its potential V_i,

$$\ddot{\phi}_i + 3H\phi_i = -\partial_i \phi_i , \tag{35.31}$$

but Hubble friction coming from all of them, since the Friedmann equation in the inflationary regime (negligible kinetic energy) is

$$3H^2 M_{\rm Pl}^2 \simeq \sum_{i=1}^{N} V_i . \tag{35.32}$$

The simplest case is when all the axions have the same mass, $m_i = \Lambda_i^2/f_i \equiv m$, and for field displacements small with respect to their relevant scale, $\phi_i \ll f_i$, so we can approximate the potential by a quadratic one. Then

$$V = \sum_i V_i \simeq \frac{1}{2} m^2 \sum_{i=1}^{N} \phi_i^2 \equiv \frac{1}{2} m^2 \Phi^2 , \tag{35.33}$$

so indeed the effective field is the radius in field space, Φ, enhanced by a $\sqrt{N}$ with respect to the individual displacements.

In order to have a Φ with a range larger than $10 M_{\rm Pl}$, and still have calculability and reasonable individual field displacements, we need a number N of axions larger than about 1000.

Note that the model says that the radial coordinate is different than the Cartesian system of coordinates. That is so, since the individual fields have periodicities $\phi_i \sim \phi + 2\pi f_i$, restricting their range, but the radial coordinate doesn't.

35.5 Axion Monodromy Inflation

The most popular type of model, originally proposed by McAllister, Silverstein and Westphal, is called axion monodromy inflation, developed from a model of inflation from general string monodromy by Silverstein and Westphal [67]. In the following I will follow mostly the arguments in [67].

Monodromy refers to the fact that when going around a closed loop in some parameter space, a system doesn't return to its original state, but to a modified one. An example relevant for the present case would be a helix or spiral, where upon performing a circular motion, the system also evolves in a third direction.

Indeed, in certain special string compactifications, while going around what looks like a loop in moduli space, specifically varying the axion across its "periodicity" range, in reality the potential energy V is modified, thus the system is not, after all, periodic. That leads to a substantial enhancement of the period, or field range, in effect becoming infinite, which as we saw before, was needed in order to make axionic inflation compatible with data.

All the relevant string monodromies refer to D-branes moving around in a (naive) configuration space, and returning to a modified state after a loop. The generic D-brane action is, as we saw,

$$S = S_{\rm DBI} + S_{\rm WZ} = -\frac{1}{(2\pi)^p {\alpha'}^{\frac{p+1}{2}} g_s} \int d^{p+1}\sigma \left[e^{-(\Phi-\Phi_0)} \sqrt{-\det(G_{MN} + B_{MN}) \partial_a X^M \partial_b X^N} + \left(\sum_p C^{(p)} \right) e^{-B} e^{2\pi\alpha' F} \right] . \tag{35.34}$$

The simplest case to consider is of a D5-brane in type IIB string theory, wrapped on a 2-cycle Σ_2, and the axion is the B-field wrapped on it, $b = \frac{1}{2\pi\alpha'} \int_{\Sigma_2} B_2$. Then the relevant term in the wrapped D5-brane action is

$$S = \frac{1}{(2\pi)^5 g_s (\alpha')^3} \int_{M_4 \times \Sigma_2} d^6\sigma \sqrt{-\det G_{ab} + B_{ab}} \, , \tag{35.35}$$

where G_{ab} and B_{ab} are the induced metric and B-field. Since we have, for a compact space with $g_{11} = g_{22} = l^2$,

$$\det(G_{ab} + B_{ab}) = \det_{M_4} g \cdot \det \begin{pmatrix} g_{11} & B_{12} \\ -B_{12} & g_{22} \end{pmatrix} = \det_{M_4} g \cdot ((g_{11})^2 - (B_{12})^2) = \det_{M_4} g \cdot (l^4 - b^2 \alpha'^2) \, , \tag{35.36}$$

the term in the action becomes a potential for the axion b,

$$S = \int d^4x \sqrt{-g} \frac{1}{(2\pi)^5 g_s (\alpha')^3} \sqrt{l^4 - b^2\alpha'^2} \equiv \int d^4x \sqrt{-g} V(b). \tag{35.37}$$

At large b, this becomes

$$V(b) \simeq \frac{1}{(2\pi)^6 g_s (\alpha')^2} b. \tag{35.38}$$

For the axion b, we have seen that the kinetic term is $-\frac{1}{2} f^2 (\partial_\mu b)^2$, and then in terms of the canonically normalized scalar $\phi = fb$, we have the action

$$S = \int d^4x \sqrt{-g} \left[-\frac{1}{2} (\partial_\mu \phi)^2 - \mu^3 \phi + V_{\rm per}(\phi) \right] , \quad \mu^3 \equiv \frac{1}{f} \frac{1}{(2\pi)^6 g_s (\alpha')^2} , \tag{35.39}$$

where $V_{\rm per}(\phi)$ is the $1 - \cos(\phi/f)$ potential for the axion, which can be neglected for the needed large b.

This action is the action for a chaotic inflation model with linear potential, but protected against corrections for super-Planckian displacements by the shift symmetry, which is just mildly broken by the new, linear, potential term.

Another possibility for inflaton monodromy from string theory can appear for D-branes wrapped on cycles in some special manifolds, where the position of the D-brane on the cycle is just the inflaton itself.

Such a manifold is the Nil 3-manifold with metric

$$\frac{ds^2}{\alpha'} = L_{u_1}^2 du_1^2 + L_{u_2}^2 du_2^2 + L_x^2 (dx' + M u_1 du_2)^2 , \tag{35.40}$$

compactified by identification under the shifts

$$\begin{aligned} &(x, u_1, u_2) \to (x + 1, u_1, u_2) \\ &(x, u_1, u_2) \to \left(x - \frac{M}{2} u_2, u_1 + 1, u_2 \right) \\ &(x, u_1, u_2) \to \left(x + \frac{M}{2} u_1, u_1, u_2 + 1 \right) . \end{aligned} \tag{35.41}$$

For a D4-brane in type IIA string theory wrapped on u_2 and moving in u_1 in the above manifold, the DBI term in the D4-brane action gives

$$S = -\frac{1}{(2\pi)^4 g_s \alpha'^{5/2}} \int_{M_4 \times u_2} d^5\sigma \sqrt{-g_4} \sqrt{g_{u_2 u_2} (1 - \alpha' g_{u_1 u_1} \dot{u}_1^2)}. \tag{35.42}$$

Doing trivially the integral over u_1 (with period 1), since nothing depends on it, and defining $L^3 \equiv L_{u_1} L_{u_2} L_x$ and L_u by $L^3 \equiv L_u^2 L_x$ and then $\beta = L_{u_2}/L_{u_1}$, we obtain a 4 dimensional action

$$S = -\frac{1}{(2\pi)^4 g_s \alpha'^{5/2}} \int_{M_4} d^4x \sqrt{-g} \sqrt{\beta L_u^2 + L_x^2 M^2 u_1^2} \sqrt{1 - \alpha' \frac{L_u^2}{\beta} \dot{u}_1^2}. \tag{35.43}$$

We see that again we obtain (for $\dot{u}_1 = 0$) a potential term $V(u_1)$ of the form $\sqrt{K^2 + u_1^2}$ as before, also going to the linear term u_1 at infinity, but now the presence of the nontrivial kinetic term for u_1 means that in terms of the canonical scalar ϕ we obtain

$$S = \int d^4x\sqrt{-g_4}\left(\frac{1}{2}\dot{\phi}_1^2 - \frac{\sqrt{\beta}L_u}{(2\pi)^4 g_s \alpha'^2}\sqrt{1 + \frac{M^2}{\beta}\frac{L_x^2}{L_u^2}u_1^2(\phi)}\right). \tag{35.44}$$

At very large u_1, we see from the above action that the first square root becomes $\propto u_1$, which means that both the potential becomes linear in u_1, and the kinetic term becomes $\propto u_1\dot{u}_1^2$, leading to a canonical scalar $\phi \propto u_1^{3/2}$, and a potential

$$V(\phi) \propto u_1 \propto \phi^{2/3}. \tag{35.45}$$

Conditions on Axion Monodromy Models

As with all string inflation models, axion monodromy has plenty of constraints to be satisfied, and it is very difficult to satisfy them all.

For one, we need as always when we have D-branes, to consider the consistency of the compactification.

Indeed, first *"tadpole cancellation,"* which amounts to having a vanishing total D-brane charge on the compact manifold (otherwise the flux has nowhere to go on the compact manifold) imposes some restrictions on possible models. As usual, this can be fixed by having an anti-D-brane as some other point in the compactification, or by adding orbifold fixed points.

Second, we need to *stabilize the moduli,* which as we saw in general, can introduce terms that spoil the flatness of the inflation potential. If the shift symmetry is not badly broken, it is now easier to satisfy this constraint than in general string inflation models, but we still have generically a problem.

For instance, if we consider the same b modulus as above, and a single Kähler modulus T, with superpotential and Kähler potential

$$\begin{aligned} W &= W_0 + Ae^{-T} \\ K &= -3\ln(T + \bar{T} + \gamma b^2)\,, \end{aligned} \tag{35.46}$$

then the holomorphic volume T is stabilized, but we obtain terms in the scalar potential that come from wrapped branes, and therefore depend on the "physical volume" $V \propto (T + \bar{T} + \gamma b^2)^{3/2}$. That means that the potential depends on both T and b, so any flat direction is lifted.

However, for axions coming from R-R fields, like the S-dual of the B field, the C_2 of type IIB string theory leading to a c axion, the problems are aleviated, as there is no tree level c term in the Kähler potential K.

There are nonperturbative corrections to the superpotential, coming from Euclidean strings and branes wrapping cycles, as well as nonperturbative gauge

theory corrections on branes (like gaugino condensation on D7-branes). These need to be taken into consideration, and to make sure nothing spoils the potential we want.

We also need to consider the *backreaction* of the branes on the geometry: in most we cannot consistently treat them in the probe approximation.

35.6 Dante's Inferno

A model that fixes some of the problems above is a variant of axion monodromy with two axions, called "Dante's Inferno". In effective field theory, we consider two axions r and θ, with axion decay constants that satisfy $f_r < f_\theta$ and

$$\begin{aligned} S &= \int d^4x\sqrt{-g}\left[-\frac{1}{2}(\partial_\mu r)^2 - \frac{1}{2}(\partial_\mu\theta)^2 - V(r,\theta)\right] \\ V(r,\theta) &= W(r) + \Lambda^4\left[1 - \cos\left(\frac{r}{f_r} - \frac{\theta}{f_\theta}\right)\right]. \end{aligned} \tag{35.47}$$

We see that the names r and θ are related to the fact that only r has a term that breaks the shift symmetry of the potential, $W(r)$, though the metric in field space is $ds^2 = dr^2 + d\theta^2$, so it is not really a "radius". The specific case studied is $W(r) = \frac{1}{2}m^2r^2$. However, if we represent (r,θ) as cylindrical coordinates, then $V(r,\theta)$ looks like a spiral with edges for each spire, which brings to mind the description of the Inferno put forward by Dante, hence the name.

If we choose parameters that satisfy

$$f_r \ll f_\theta \ll M_{\rm Pl}\ , \quad \Lambda^4 \gg f m^2 r_{\rm in}\ , \tag{35.48}$$

where $f \ll r_\in < M_{\rm Pl}$ for the initial condition in r, then the excitation in the $\tilde{r}$ direction, appearing in the cosine potential, is very massive, and $\tilde{r}$ is effetively fixed. The remaining coordinate $\tilde{\theta}$ is the inflaton. The field redefinition (rotation) is

$$\begin{pmatrix}\tilde{r}\\ \tilde{\theta}\end{pmatrix} = \begin{pmatrix}\cos\xi & \sin\xi\\ -\sin\xi & \cos\xi\end{pmatrix}\begin{pmatrix}r\\ \theta\end{pmatrix}\ , \tag{35.49}$$

where

$$\sin\xi = \frac{f_r}{\sqrt{f_r^2+f_\theta^2}}\ , \quad \cos\xi = \frac{f_\theta}{\sqrt{f_r^2+f_\theta^2}}\ , \quad f = \frac{f_r f_\theta}{\sqrt{f_r^2+f_\theta^2}}. \tag{35.50}$$

In our limit on the parameters, we obtain that, since $\tilde{r}$ is almost fixed, the effective field is

$$\phi_{\rm eff} = \tilde{\theta} \simeq \theta, \tag{35.51}$$

its potential is

$$V_{\rm eff}(\phi_{\rm eff}) = \frac{1}{2} m_{\rm eff}^2 \phi_{\rm eff}^2 \ , \tag{35.52}$$

and the effective mass of the field is given by

$$m_{\rm eff}^2 = m^2 \sin^2 \xi \simeq m^2 \frac{f_r^2}{f_\theta^2}. \tag{35.53}$$

That means that we obtain a classic chaotic inflation model (other power laws are also possible), but the advantages are:

- the effective mass is suppressed compared with the mass in the Lagrangian: $m_{\rm eff} \simeq m f_r / f_\theta \ll m$.
- the Lyth bound is modified: indeed, the Lyth bound about transplanckianity needed for nontrivial r ratio of tensors to scalars refers to the effective field $\phi_{\rm eff} \simeq \theta$, which needs to be $\gg M_{\rm Pl}$. But the θ direction is periodic, which helps it avoid large corrections. On the other hand the range in r, which is the field with a mass term, is

$$\Delta r \simeq \frac{f_r}{\sqrt{f_r^2 + f_\theta^2}} \Delta\phi_{\rm eff} \simeq \frac{f_r}{f_\theta} \Delta\phi_{\rm eff} \ll \Delta\phi_{\rm eff} \ , \tag{35.54}$$

  and can be made subplanckian.

More importantly, the model can be embedded into string theory as follows. Consider an Euclidean D1-brane instanton, wrapping a 2-cycle Σ that is a linear combination of the cycles Σ_r and Σ_θ that correspond to the axions r and θ (the axions are fields integrated on them). Then we obtain a term in the potential of the required form,

$$\cos\left(\alpha \frac{r}{f_r} - \beta \frac{\theta}{f_\theta}\right). \tag{35.55}$$

One also finds that it is not hard to obtain the needed hierarchy $f_r / f_\theta \ll 1$.

35.7 Phenomenology of Axion Monodromy Inflation

The axion monodromy potential is

$$V = V_0(\phi) + \Lambda^4 \cos \frac{\phi}{f} \ , \tag{35.56}$$

and one needs the second term to be subdominant, but not small,

$$b_* \equiv \frac{\Lambda^4}{V_0'(\phi_*) f} < 1 \,, \tag{35.57}$$

where ϕ_* is the value of the inflaton when the pivot scale k_* exits the horizon.

There are many models of axion inflation from string theory, but there are a few common signatures for them. Among those, we have:

- because of the modulated potential for the axion, one finds a modulation in the amplitude of the power spectrum $\Delta^2_R(k)$, to leading order in b_*,

$$\Delta^2_R(k) = \Delta^2_R(k_*) \left(\frac{k}{k_*}\right)^{n_s - 1} \left[1 + A \cos\left(\frac{\phi_k}{f}\right)\right] \,, \tag{35.58}$$

  where $A \propto b_*$.
- the non-Gaussianity is oscillating, also because of the oscillating potential, and its possible resonance means that the constraints on the parameters from observations are looser than for usual models.
- the inflation is coupled to gauge fields, which results in gauge field production.
- the produced gauge fields have a stress tensor that can source gravitational waves, which are *chiral*, i.e., of only one chirality: the one that is unstable for the gauge field.

Important Concepts to Remember

- The vacuum state in the presence of instantons is parametrized by a phase θ, mixing winding numbers, and which can be represented as a parameter in the Lagrangian, multiplying the instanton action $F * F$.
- Axions can explain the smallness of θ: the anomaly equation in the presence of instantons means a cosine potential for θ, minimized at $\theta = 0$; $\theta(x)$ is called the axion field.
- Axions in string theory can appear as p-form fields integrated over cycles, or 4-dimensional duals of p-form fields, and are classically shift symmetric.
- At the quantum level, the string theory axions can get nonperturbative corrections from worldsheet, or D-brane, instantons.
- For a good axionic string inflation model, we need $f > 10 M_{\rm Pl}$, which is not possible for a generic string axion.
- We can have a fine-tuned "racetrack" with two axions, where an effective remaining axion has a very large decay constant f.
- We can have N-flation for the axions, where the effective field is the radial coordinate of N identical fields, and $\Phi \simeq \sqrt{N}|\phi|$.
- Axion monodromy means that the axion gets an extra $V_0(\phi)$, besides the axionic potential.
- For the axion monodromy model of a D5-brane in type IIB, with a B-field on Σ_2 as the axion, we have at large b a linear potential, leading to a chaotic inflation model.

- For a simple string monodromy model, of D4-brane moving on a cycle in a Nil 3-manifold, the inflaton potential is $V \propto \phi^{2/3}$.
- Dante's Inferno is a 2-axion monodromy model, with one field having a steep potential like a mass term, and a linear combination having the consine potential, and a hierarchy on the decay constants $f_r \ll f_\theta$.
- Axion monodromy produces an oscillating power spectrum, oscillating non-Gaussianity, gauge fields, and chiral gravitational waves.

Further reading: For more about axions, see [1], Chap. 10. For more about axions in string theory, and axion monodromy models, see [31], Sects. 3.2 and 5.4. The original axion monodromy scenario was defined in [66], following the model of [67] for inflation from monodromy in string theory, and the "Dante's Inferno" scenario was defined in [68].

Exercises

(1) Show the details for the calculation of the DBI action of the D4-brane on the Nil 3-manifold (35.40) leading to the action (35.44).
(2) Show explicitly that for

$$\begin{aligned} W &= W_0 + Ae^{-T} \\ K &= -3\ln(T + \bar{T} + \gamma b^2)\ , \end{aligned} \tag{35.59}$$

the holomorphic volume T is stabilized.
(3) Calculate the CMBR parameters r and n_s for the Dante's Inferno model, as a function of its parameters.
(4) Consider the D-brane action for $p = 5$. Does the WZ action affect the potential for the axion ϕ for any constant $C^{(p)}$?

Chapter 36
Fuzzy Dark Matter from String Theory

In this chapter, we will examine an alternative to the usual ΛCDM model, where the cold dark matter is replaced by a "fuzzy dark matter", and we will see that it appears more naturally in string theory. In fact, it is based on the axionic models from the previous chapter.

36.1 Fuzzy Dark Matter from String Theory

We have seen that a priori, the two most important relevant cases of dark matter models were cold dark matter, the relatively heavy, nonrelativistic matter that can only interact with itself but not others, and hot dark matter, which meant thermalized dark matter, in practice neutrinos, which are still relativistic (with a temperature comparable to the one of the CMBR), and are thermalized (because of their early times interaction with the other particles). Neutrinos are pretty much the only possibilities for hot dark matter, and have masses of less than 10^{-2} eV. As we saw in Chap. 3, the hot dark matter scenario is actually ruled out, though a small hot component to CDM ("warm dark matter") is still possible.

But there is a third possibility, one that is intriguing, and relates to the axions of the previous chapter. We can have dark matter that is extremely light, with masses of the order of $m \sim 10^{-22}$ eV, specifically a boson with de Broglie wavelength of $\lambda \sim 1kpc$, yet it is not thermal (extremely cold). That means that the boson, despite being much lighter than the neutrinos, is nonrelativistic, and is quantum on scales smaller than $1kpc$. The reason this scenario is possible is that the boson is not thermalized like the neutrinos.

The boson one will consider is an axion a, with action

$$S = \int d^4x\sqrt{-g}\left[-\frac{1}{2}f^2(\partial_\mu a)^2 - \Lambda^4(1-\cos a)\right]. \tag{36.1}$$

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_36

We want the mass m to be

$$m = \frac{\Lambda^2}{f} \sim 10^{-22} - 10^{-21}\,\text{eV}. \tag{36.2}$$

From string theory, the generic axion decay constant is below the Planck scale, and above the grand unification (GUT) scale, so

$$10^{18}\,\text{GeV} \gtrsim f \gtrsim 10^{16}\,\text{GeV}\,, \tag{36.3}$$

though it is worth noting that most models from the previous chapter obtained a larger f for the effective axion. The confining scale Λ is obtained from instanton effects, and can in principle be suppressed by susy effects, with scale Λ_S, and we have

$$\Lambda^4 \sim M_{\text{Pl}}^2 \Lambda_S^2 e^{-S_{\text{inst}}}\,, \tag{36.4}$$

where S_{inst} is the instanton action, and Λ_S has, in susy breaking scenarios (or without susy at all), the value

$$10^{18}\,\text{GeV} \gtrsim \Lambda \gtrsim 10^{4}\,\text{GeV}. \tag{36.5}$$

To obtain a mass of $m \sim 10^{-22}\,\text{eV}$, if $\Lambda = 10^4\,\text{GeV}$ we need $S_{\text{inst}} = 165$, and if $\Lambda = 10^{18}\,\text{GeV}$, we need $S_{\text{inst}} = 230$. But from the Standard Model, the instanton action is $S_{\text{inst}} = 8\pi^2/g^2 \equiv 2\pi/\alpha_G$, where $\alpha_G = g^2/(4\pi)$ is the Standard Model gauge coupling at the GUT scale $\sim 10^{16}\,\text{GeV}$. For $\alpha_G = 1/20$, we get $S_{\text{inst}} = 126$, and for $\alpha_G = 1/30$, we get $S_{\text{inst}} = 188$, so there is overlap with the desired region.

36.2 Axion Dynamics in the Expanding Universe

Having seen what the basic set-up is, and that it can be derived from string theory, we turn to describing its consequences.

The first thing to do is to set up how the axion evolves from the beginning of the Universe until today.

One assumes that the axion starts off as a constant field with some random initial value a_0. Then we evolve it in FLRW Universe $ds^2 = -c^2dt^2 + R^2(t)d\vec{x}^2$ as a time dependent field, $a = a(t)$. The KG equation of motion for the axion is then

$$\ddot{a} + 3H\dot{a} + m^2 \sin a = 0. \tag{36.6}$$

The solution is approximately constant for $H \gtrsim m$, and for $H \lesssim m$, it oscillates with angular frequency m, damped as $R(t)^{-3/2}$ due to the usual Hubble friction term.

The transition between the two regimes happens at a temperature T_0 such that (since we assume that the Universe is radiation dominated at early times) $\rho \simeq \rho_{\text{rad}} \sim T_0^4 \sim H^2 M_{\text{Pl}}^2 \sim m^2 M_{\text{Pl}}^2$, so

$$T_0 \sim \sqrt{m M_{\rm Pl}} \sim 500\,{\rm eV}\ , \tag{36.7}$$

or a redshift of about $z \sim 2 \times 10^6$, which is after nucleosynthesis. Note that matter domination starts at about $T_1 \sim 1\,{\rm eV}$, so this is still after T_0.

Since the axion is about constant until $T_0 \sim 500\,{\rm eV}$, its potential energy (there is no kinetic energy since $\vec{\nabla} a = 0$ and $\dot{a} \simeq 0$ on the solution) $V \sim \Lambda^4$ acts as a component of dark energy ("cosmological constant"), being constant.

But after $T_0 \sim 500\,{\rm eV}$, the axion starts to oscillate (damped by Hubble friction), and it thus behaves like a condensate of bosons of zero spatial momentum, with energy density dropping like $R(t)^{-3}$, as for any nonrelativistic matter. It thus behaves like an extremely cold nonrelativistic dark matter (of negligible temperature).

One can also show that the axion self-interaction (due to the cosine potential) becomes extremely quickly negligible, in fact already at $T = 10^{-5/3} T_0$.

We can fix the the axion decay constant by noting that after T_0 (when radiation has $\rho_{\rm rad} \sim T_0^4$ and (fuzzy dark) matter has $\rho_m \sim \Lambda^4 = f^2 m^2$) and until T_1, when $\rho_m \sim \rho_{\rm rad}$, we have $\rho_m / \rho_{\rm rad} \propto T$, so

$$\left.\frac{\rho_m}{\rho_{\rm rad}}\right|_{eq.} \sim \frac{\Lambda^4}{T_0^4} \times \frac{T_0}{T_1} \simeq 1 \Rightarrow f = \sqrt{\frac{\Lambda^2}{m}} \sim \frac{M_{\rm Pl}^{3/2} T_1^{1/2}}{m^{1/4}} \sim 0.5 \times 10^{17}\,{\rm GeV}\ , \tag{36.8}$$

which is consistent with the range required earlier.

36.3 Fuzzy Dark Matter as a Nonrelativistic Quantum System

The main applications of fuzzy dark matter are for galaxies and structures present in the Hubble volume, as in the case of cold dark matter, so we need to find out its description today. The first thing to note is that we can describe the fuzzy dark matter as a nonrelativistic quantum system.

The axion must be nonrelativistic, by our assumption. We thus write its action by reintroducing $\hbar$ and c, with the purpose of taking the nonrelativistic limit:

$$S = \int \frac{d^4x}{\hbar c^2} \sqrt{-g} \left[-\frac{1}{2} g^{\mu\nu} \partial_\mu \phi \partial_\nu \phi - \frac{1}{2} \frac{m^2 c^2}{\hbar^2} \phi^2 \right]\ , \tag{36.9}$$

where the field ϕ has dimensions of energy and $x^0 = ct$. The de Broglie wavelength is given by

$$\frac{\lambda}{2\pi} = \frac{\hbar}{m v} = 1.92 kpc \left(\frac{10^{-22}\,{\rm eV}}{m} \right) \left(\frac{10\,{\rm km/s}}{v} \right)\ , \tag{36.10}$$

which means that in order to get a de Broglie wavelength of about kpc, we need $m \sim 10^{-22}$ and $v \sim 10\,{\rm km/s}$, which indeed makes it very non-relativistic.

In the nonrelativistic limit, we must take out a phase containing the rest energy of the particle, $e^{-i\frac{E_0 t}{\hbar}} = e^{-i\frac{mc^2}{\hbar^2}}$, as well as a constant that rescales the action to the nonrelativistic action. Moreover, then the field multiplying it is complex, and we must add the complex conjugate, so

$$\phi = \sqrt{\frac{\hbar^2 c}{2m}} \left(\psi e^{-i\frac{mc^2}{\hbar}} + \psi^* e^{+i\frac{mc^2}{\hbar}} \right). \tag{36.11}$$

Substituting in the action, we obtain

$$S = \int d^4x \sqrt{-g} \left[-|g^{00}| \psi^* i\hbar \partial_t \psi + g^{rr} \frac{\hbar^2}{2m} (\vec{\nabla}\psi^*)(\vec{\nabla}\psi) + (|g^{00}| - 1) \frac{mc^2}{2} \psi^* \psi \right]. \tag{36.12}$$

Considering a background of the type of a FLRW metric with scale $R(t)$, perturbed by a Newton potential $\Phi(\vec{r}, t)$,

$$ds^2 = -\left(1 + \frac{2\Phi}{c^2}\right) c^2 dt^2 + R^2(t) \left(1 - \frac{2\Phi}{c^2}\right) d\vec{r}^2 \,, \tag{36.13}$$

and defining as usual the Hubble constant as $H = \dot{R}/R$, we find the equation of motion (varying with respect to ψ^*)

$$i\hbar \left(\dot{\psi} + \frac{3}{2} H \psi \right) = \left(-\frac{\hbar^2}{2m R^2(t)} \vec{\nabla}^2 + m\Phi \right) \psi. \tag{36.14}$$

This is just a modified Schrödinger equation, with terms coming from the Hubble expansion and the Newton potential, as expected for the nonrelativistic limit of a scalar field theory.

36.4 Superfluid Picture

However, we are in a classical situation, with small quantum fluctuations, so we must be able to view ψ as some kind of classical field. That is only possible if we view the dark matter as a superfluid. A regular fluid is a description of a quantum mechanical system at large distances (ignoring terms of higher orders in derivatives), whereas a superfluid is a quantum fluid with phase coherence: the wavefunction of the superfluid defines both the density ρ of the fluid, through the interpretation of $|\psi|^2$ as probability density, i.e., number density, or ρ/m, whereas the $U(1)$ current $\vec{j}$ defined from the phase is understood as $\rho\vec{v}$. That means that

$$\psi = \sqrt{\frac{\rho}{m}} e^{i\theta} \tag{36.15}$$

and the velocity is

$$\vec{v} = \frac{\vec{j}}{\rho} = \frac{1}{m\psi^*\psi}\frac{-i\hbar}{2R(t)}(\psi^*\vec{\nabla}\psi - \psi\vec{\nabla}\psi^*) = \frac{\hbar}{mR(t)}\vec{\nabla}\theta. \tag{36.16}$$

Because of $\vec{v} \propto \vec{\nabla}\theta$, we have zero vorticity, $\vec{\omega} \equiv \vec{\nabla} \times \vec{v} = 0$.

Replacing the above ψ in the modified Schrödinger equation, separating the equation into one for the modulus and one for the phase, then taking $\vec{\nabla}$ of the phase and replacing $\vec{\nabla}\theta$ by $\vec{v}$, we obtain two equations that look like the ones of the fluid in comoving coordinates,

$$\begin{aligned}
&\dot{\rho} + 3H\rho + \frac{1}{R}\vec{\nabla}\cdot(\rho\vec{v}) = 0 \\
&\dot{\vec{v}} + H\vec{v} + \frac{1}{R}(\vec{v}\cdot\vec{\nabla})\vec{v} = -\frac{1}{R}\vec{\nabla}\Phi + \frac{\hbar^2}{2R^3m^2}\vec{\nabla}\left(\frac{\vec{\nabla}^2\sqrt{\rho}}{\sqrt{\rho}}\right).
\end{aligned} \tag{36.17}$$

The first is like the continuity equation, and the second like the Euler equation, generalized to the expanding Universe. These are known as the Madelung equations, and we note that the only new feature is the "quantum pressure term," the last term in the Euler equation.

One can numerically simulate the Madelung equations to find specific fuzzy dark matter solutions.

36.5 Schrödinger–Poisson Equation and Spherical Soliton Solution

However, there is another way to approach solutions of the equations of motion (36.14). For many applications for galaxies, it is enough to ignore the Hubble expansion, so put $H = 0$ and $R(t) = 1$. Then we have a usual Schrödinger equation, turned into a time independent one in the usual way, by $\psi(\vec{r}, t) = e^{-iEt}\psi(\vec{r})$, so

$$-\frac{\hbar^2}{2m}\vec{\nabla}^2\psi(\vec{r}) + m\Phi\psi(\vec{r}) = E\psi(\vec{r}). \tag{36.18}$$

The new feature is that now, as before, $|\psi|^2$ is a probability density, so $m|\psi|^2$ is a mass density, that sources the Newton potential, so we also have the Poisson equation

$$\vec{\nabla}^2\Phi = 4\pi G_N m|\psi|^2. \tag{36.19}$$

For the solution defining a galaxy, we consider an isolated system, assuming that as $|\vec{r}| \to \infty$, the variables $\psi(\vec{r})$ and $\Phi(\vec{r})$ go to zero.

Considering a spherical solution, which can be chosen to be real, we obtain the equations

$$-\frac{\hbar^2}{2mr^2}\frac{d}{dr}r^2\frac{d\psi}{dr}+m\Phi\psi=E\psi$$
$$\frac{1}{r^2}\frac{d}{dr}r^2\frac{d\Phi}{dr}=4\pi G_N m\psi^2. \tag{36.20}$$

We obtain soliton solutions ordered by an energy level n (for eigenstates of the Schrödinger equation), ψ_n and Φ_n. It turns out that we can scale out dimensions from all the parameters, given an overall mass $M=\int\rho=m\int|\psi|^2$ for the soliton, and express them in terms of numerically calculable dimensionless numbers. We find that the central density, central potential, and half-mass radius are

$$\rho_c=\left(\frac{Gm^2}{\hbar^2}\right)^3 M^4\rho_n\ ,\quad \Phi_c=\left(\frac{GMm}{\hbar}\right)^2\phi_n\ ,\quad r_{1/2}=\frac{\hbar^2}{GMm^2}f_n. \tag{36.21}$$

Numerically, we see that $f_0\simeq 3.9251$ and f_n increases for $n>0$, $\rho_0=0.00440$ and ρ_n decreases for $n>0$.

36.6 Observational Consequences

First, the soliton formulas above mean that

$$r_{1/2}\geq 3.925\frac{\hbar^2}{GMm^2};\quad \rho_c\leq 0.0044\left(\frac{Gm^2}{\hbar^2}\right)^3 M^4\ , \tag{36.22}$$

so there is a minimum half-mass radius, of about $0.33kpc$ for $M=10^9M_{\rm Sun}$ and $m=10^{-22}$ eV, and a maximum central density, of about $7.05M_{\rm Sun}/pc^3$ for the same values.

There is also a maximum mass for the soliton, since the central Newton potential Φ of the soliton cannot exceed $\sim c^2$, which from (36.21) gives about $\hbar c/(G_N m)$. More precisely, one finds

$$M_{\rm max}=0.633\frac{\hbar c}{G_N m}\ , \tag{36.23}$$

which is about $8.46\times 10^{11}M_{\rm Sun}$ for $m=10^{-22}$ eV.

For distances large compared to the de Broglie wavelength, which as we saw, was about kpc order, fuzzy dark matter behaves like ordinary dark matter, since it looks classical. So for any scale larger than galactic, we cannot see a difference.

Differences occur on smaller scales. The minimum half-mass radius and maximum central density are examples.

The soliton core is expected to be surrounded by a virialized halo with the usual (CDM) Navarro-Frenk-White (NFW) profile, since those distances are larger than the de Broglie wavelength. The maximum mass above is only the mass of the soliton core, there are FDM halos with much larger mass.

From simulations, one finds a minimum mass for dark matter halos of the order of $10^7 M_{\rm Sun}(m/10^{-22}\,{\rm eV})^{-3/2}$.

Dynamical friction for clusters moving in dwarf galaxies is reduced considerably in FDM, due to various effects, including the quantum effects on small scales (smaller than the de Broglie wavelength) of the superfluid. That means that the timescale for relaxation is greatly increased, as observed experimentally in the globular clusters of the Fornax dwarf galaxy, a satellite galaxy of our own Milky Way galaxy.

The only possible contradiction with observations of the fuzzy dark matter scenario is from the Lyman-α absorbtion spectrum in galaxies, but there are many unknown effects, and it is not clear if the constraint, which seems to favor a mass m of about $10-20\times 10^{-22}\,{\rm eV}$ rather than the $1-10\times 10^{-22}\,{\rm eV}$ range favoured by the rest of observations, is correct.

Important Concepts to Remember

- Fuzzy dark matter is an alternative to cold dark matter, with an ultra-light boson with a mass $m\sim 10^{-22}\,{\rm eV}$, and a de Broglie wavelength of about $1kpc$, which means it is classical on larger scales.
- The boson is ultra-cold, so is nonrelativistic despite its extremely small mass.
- The best candidate is an axion, specifically a string theory axion with a decay constant f in between the Planck and GUT scales, and an energy density scale larger than $10^4\,{\rm GeV}$.
- The axion behaves like a dark energy component, being a constant, until $T_0\sim\sqrt{mM_{\rm Pl}}\sim 500\,{\rm eV}$, and as regular dark matter afterwards.
- FDM behaves like a nonrelativistic quantum system, satisfying a modified Schrödinger equation for the wavefunction, with a Hubble friction and a Newton potential term.
- FDM can be also recast as a superfluid satisfying Madelung equations, with a quantum pressure term.
- On galaxy scales, the Schrödinger–Poisson system admits spherical scaling soliton solutions.
- The soliton picture results in a maximum central core density, a minimum half-mass radius, a maximum core mass, and an NFW halo outside.

Further reading: The fuzzy dark matter scenario with string theory applications was defined in [69].

Exercises

(1) Show that the free massive scalar action (36.9) gives in the nonrelativistic limit (36.12), and that its equations of motion in the FLRW metric perturbed by the Newton potential, (36.13), are given by the modified Schrödinger equation (36.14).
(2) Show that the modified Schrödinger equation (36.14) leads to the Madelung equations (36.17).

(3) Prove the scaling relations (36.21) for the radial soliton solutions of the Schrödinger–Poisson equations.
(4) Find the solution for the equation of motion of the axion field (KG) in the expanding Universe, (36.6).

Chapter 37
Holographic Cosmology

In this chapter we will consider holographic approaches to cosmology, that is, to take advantage of the duality to a weakly coupled (therefore calculable) field theory provided by the AdS/CFT correspondence and its generalizations, gauge/gravity duality, in order to calculate the behaviour near the strongly coupled gravity near the cosmological singularity.

There have been various attempts at using holography in this way, with various degrees of success. Here I will focus on two examples, which provide more concrete observational consequences, though they both come with caveats.

In one approach, started by McFadden and Skenderis, one starts with a "domain wall" spacetime, which is an Euclidean version of the FLRW cosmology in the spatially flat ($k = 0$) case. One then constructs a *holographic phenomenological approach*, by calculating the results of a generic super-renormalizable quantum field theory (at the 3-dimensional boundary of the domain wall spacetime) momentum space 2-point function and, after analytical continuation to Lorentzian signature, compare the resulting power spectra against observations.

In another approach, based on some exact (*top down holography*) type IIB supergravity cosmological solutions, one calculates the transition of cosmological perturbations through a (strongly coupled) cosmological singularity, by doing a calculation in the (weakly coupled) gauge theory, assumed to be $\mathcal{N} = 4$ SYM with a time-dependent coupling constant g_{YM}, and then mapping to the dual cosmology.

37.1 Generic Holographic Cosmology Map

The generic idea about using holography to calculate correlation functions in cosmology is due to the seminal paper of Maldacena from 2002, which calculated for the first time inflationary non-Gaussianity in the 3-point functions. Maldacena proposed that holography in an asymptotically de Sitter background should use the

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8_37

correspondence between the wavefunction of the Universe $\Psi[h_{ij}, \phi]$, with a given spatial 3-metric h_{ij} and scalar ϕ, and the dual field theory partition function, with the partition function $Z_{\rm QFT}[h_{ij}, \phi]$ depending on sources h_{ij} and ϕ for the operators,

$$\Psi[h_{ij}, \phi] = Z_{\rm QFT}[h_{ij}, \phi]. \tag{37.1}$$

The correspondence is analytically continued from (asymptotically) AdS space. The fields in the cosmology side start in the Bunch–Davies vacuum, corresponding to the analytical continuation from Euclidean signature.

The the stress tensor T_{ij} on the boundary is expressed in terms of the wavefunction Ψ in the bulk as

$$T_{ij}(x) = \left.\frac{\delta Z_{\rm QFT}[h_{ij}, \phi]}{\sqrt{h}\delta h^{ij}(x)}\right|_{\delta h_{ij}=0,\phi=0} = \left.\frac{\delta \Psi[h_{ij}, \phi]}{\sqrt{h}\delta h^{ij}(x)}\right|_{\delta h_{ij}=0,\phi=0}. \tag{37.2}$$

This is understood as an operator statement, and valid inside correlation functions. Analogously, we have for the boundary operator $\mathcal{O}$ coupling with ϕ a variation of the usual AdS/CFT relation,

$$\langle \mathcal{O}(x_1) \dots \mathcal{O}(x_n)\rangle = \left.\frac{\delta^n Z_{\rm QFT}[h_{ij}, \phi]}{\delta\phi(x_1)\dots\delta\phi(x_n)}\right|_{\delta h_{ij}=\phi=0} = \left.\frac{\delta^n \Psi[h_{ij}, \phi]}{\delta\phi(x_1)\dots\delta\phi(x_n)}\right|_{\delta h_{ij}=\phi=0}. \tag{37.3}$$

This then integrates to the wavefunction of the Universe

$$\Psi[\phi] = \exp\left[\sum_{n=2}^{\infty}\frac{1}{n!}\int dx_1 \dots \int dx_n \langle \mathcal{O}(x_1)\dots\mathcal{O}(x_n)\rangle\right]. \tag{37.4}$$

and similarly when we include h_{ij}, dual to T_{ij} correlators.

The wavefunction $\Psi[h_{ij}, \phi]$ can be computed in the cosmology as $\Psi \sim e^{iS_{\rm cl}}$, where $S_{\rm cl}$ is the classical action on the on-shell fields with the given boundary values. But then correlation functions of h_{ij} and ϕ are evaluated as usual in quantum mechanics by integrating with the probability $|\Psi|^2$, e.g.

$$\langle h_{ij}(x_1)h_{kl}(x_2)\rangle = \int Dh_{mn}|\Psi[h_{mn}, \phi]|^2 h_{ij}(x_1)h_{kl}(x_2). \tag{37.5}$$

Since both sides of the equality are real, only the real part of $iS_{\rm cl}$ contributes, but we obtain a crucial factor of 2 from the square of the wavefunction:

$$\langle h_{ij}(x_1)h_{kl}(x_2)\rangle = \int Dh_{mn} e^{-2{\rm Im}S_{\rm cl}} h_{ij}(x_1)h_{kl}(x_2). \tag{37.6}$$

If the gravity theory is strongly coupled, this correlation function might be difficult to calculate. Otherwise, for a theory approximately in de Sitter space, these correlators for superhorizon distances, which are what is needed for inflationary CMB

calculations, can be evaluated (and were evaluated by Maldacena). But in a different background, it might be hard to calculate.

Note that this is the correlation function that we want, which is a bit different than the correlation functions in usual AdS/CFT, which are correlation functions of the operators $\mathcal{O}(x)$ on the boundary field theory, and not of the sources $\phi(x)$ and $h_{ij}(x)$ associated with them. If we want to use holography to calculate the correlators of ϕ and h, we need to find a way to relate them to the correlators of $\mathcal{O}$ and T_{ij} in the quantum field theory.

37.2 Domain Wall/Cosmology Correspondence

The first holographic approach is based on the "domain wall/cosmology correspondence," which means a certain Wick rotation in the context of holographic models.

The "domain wall" spacetime is really a FLRW cosmological model, analytically continued to Euclidean signature. Together, the two signature solutions are written as

$$\begin{aligned} ds^2 &= \eta dz^2 + a^2(z)[\delta_{ij} + h_{ij}(z,\vec{x})]dx^i dx^j \\ \Phi &= \phi(z) + \delta\phi(z,\vec{x}). \end{aligned} \tag{37.7}$$

Here $\eta = -1$ is FLRW cosmology, and $\eta = +1$ is the domain wall, obtained by the Wick rotation $t = -iz$. h_{ij} is a gravitational fluctuation, $\delta\phi$ a dilaton fluctuation, and $a(z)$ the scale factor. The Euclidean vacuum becomes the standard Bunch-Davies vacuum when Wick rotating to the cosmology. Both signature solutions come from the same action, just with a different overall sign,

$$S = \frac{\eta}{2\kappa^2}\int d^4x\sqrt{|g|}[-R + (\partial_\mu\phi)^2 + 2\kappa^2 V(\Phi)]. \tag{37.8}$$

If $\phi(z)$ is monotonic, we can invert it to $z(\phi)$, and thus define a "fake superpotential" $W(\phi)$ by

$$H(z) = -\frac{1}{2}W(\phi(z)). \tag{37.9}$$

Then the equations of motion of the gravity plus scalar system are

$$\frac{\dot{a}}{a} = -\frac{1}{2}W(\phi(z))\ , \quad \dot{\phi} = W'\ , \quad 2\eta\kappa^2 V = W'^2 - \frac{3}{2}W^2. \tag{37.10}$$

We can add perturbations on top of the background satisfying the above conditions, and as usual consider

$$h_{ij} = -2\psi(z,\vec{x})\delta_{ij}\ , \tag{37.11}$$

go to comoving momentum space $\vec{q}$, and construct the usual gauge invariant variable

$$\zeta = \psi + \frac{H}{\dot{\phi}}\delta\phi \tag{37.12}$$

and transverse traceless perturbation γ_{ij}, and the slow roll parameter

$$\epsilon_H = -\frac{\dot{H}}{H^2} = 2\frac{W'^2}{W^2}. \tag{37.13}$$

Then the equations of motion become as usual

$$\begin{aligned} &\ddot{\zeta} + \left(3H + \frac{\dot{\epsilon}_H}{\epsilon_H}\right)\dot{\zeta} - \eta\frac{q^2}{a^2}\zeta = 0 \\ &\ddot{\gamma}_{ij} + 3H\dot{\gamma}_{ij} - \eta\frac{q^2}{a^2}\gamma_{ij}. \end{aligned} \tag{37.14}$$

The analytical continuation from Euclidean to Lorentzian signature is

$$\bar{\kappa}^2 = -\kappa^2\ , \quad \bar{q} = -iq. \tag{37.15}$$

The scalar and tensor (superhorizon) power spectra are momentum space two-point functions of the perturbations ζ and γ_{ij} that go to a constant at late time values,

$$\begin{aligned} \Delta^2_S(q) &= \frac{q^3}{2\pi_2^2}\langle\zeta(q)\zeta(-q)\rangle = \frac{q^3}{2\pi^2}|\zeta_{q(0)}|^2 \\ \Delta^2_T(q) &= \frac{q^3}{2\pi^2}\langle\gamma_{ij}(q)\gamma_{ij}(-q)\rangle = \frac{2q^3}{\pi^2}|\gamma_{q(0)}|^2\ , \end{aligned} \tag{37.16}$$

where $\zeta_{q(0)}$ and $\gamma_{q(0)}$ are the late time values of the perturbations, which have at early times the Bunch-Davies vacuum condition $\zeta_q, \gamma_q \sim e^{-iq\tau}$ for conformal time $\tau \to -\infty$.

Define the canonical conjugate momenta to ζ_q and γ_q,

$$\Pi_q^{(\zeta)} = \frac{2}{\kappa^2}\epsilon_H a^3\dot{\zeta}_q\ , \quad \Pi_q^{(\gamma)} = \frac{1}{4\kappa^2}a^3\dot{\gamma}_q\ , \tag{37.17}$$

and impose (operatorial) canonical commutation conditions for them,

$$\begin{aligned} &\zeta_q\Pi_q^{(\zeta)*} - \Pi_q^{(\zeta)}\zeta_q^* = i \\ &2(\gamma_q\Pi_q^{(\gamma)*} - \Pi_q^{(\gamma)}\gamma_q^*) = i. \end{aligned} \tag{37.18}$$

The specific functions for ζ_q and γ_q imply a given $\dot{\zeta}_q, \dot{\gamma}_q$, therefore a given $\Pi_q^{(\zeta)}$ and $\Pi_q^{(\gamma)}$, which can be described by saying that there is a response to ζ_q and γ_q, defined by response functions Ω and E,

$$\Pi_q^{(\zeta)} = \Omega\zeta_q \ , \quad \Pi_q^{(\gamma)} = E\gamma_q . \tag{37.19}$$

Substituting this functional form in the canonical commutation relations above, the real parts cancel, and we obtain the conditions

$$|\zeta_q|^2 = -\frac{1}{2\mathrm{Im}\Omega} \ , \quad |\gamma_q|^2 = -\frac{1}{4\mathrm{Im}E} . \tag{37.20}$$

Substituting in the superhorizon power spectra, we obtain

$$\begin{aligned} \Delta_S^2(q) &= \frac{-q^3}{4\pi^2 \mathrm{Im}\Omega_{(0)}(q)} \ , \\ \Delta_T^2(q) &= \frac{-q^3}{2\pi^2 \mathrm{Im}E_{(0)}(q)} . \end{aligned} \tag{37.21}$$

In the domain wall spacetime, i.e., in Euclidean signature, we can define the same variables, via the analytical continuation

$$\bar{\Omega}(-iq) = \Omega(q) \ , \quad \bar{E}(-iq) = E(q) \ , \tag{37.22}$$

leading to the definitions

$$\bar{\Pi}_{\bar{q}}^{(\zeta)} = -\bar{\Omega}\zeta_{\bar{q}} \ , \quad \bar{\Pi}_q^{(\gamma)} = -\bar{E}\gamma_{\bar{q}} . \tag{37.23}$$

37.3 Holographic Calculation

To calculate holographically the power spectra, we must calculate $\bar{\Omega}_{(0)}$ and $\bar{E}_{(0)}$ holographically, and then make the analytical continuation to cosmology.

For asymptotically AdS or dS domain walls, the holographic prescription is understood. In particular, for the dS case relevant for inflation, this was done by Maldacena, as reviewed at the beginning of the chapter.

We can then adopt the same logic as there for the relevant case here, which is of a power law corresponding to FLRW, at least asymptotically

$$a(z) \sim (z/z_0)^n \ , \quad \phi \sim \sqrt{2n}\log(z/z_0) . \tag{37.24}$$

One calculates holographically, as described at the begining of the chapter, the two-point function of energy-momentum tensors on the boundary, decomposed in general as

$$\langle T_{ij}(\bar{q})T_{kl}(-\bar{q})\rangle = A(\bar{q})\Pi_{ijkl} + B(\bar{q})\pi_{ij}\pi_{kl} \ , \tag{37.25}$$

where

$$\Pi_{ijkl} = \pi_{i(k}\pi_{l)j} - \frac{1}{2}\pi_{ij}\pi_{kl} \ , \quad \pi_{ij} = \delta_{ij} - \frac{\bar{q}_i \bar{q}_j}{\bar{q}^2} \tag{37.26}$$

are the 4-index transverse traceless projection operator, and the 2-index transverse projection operator (a 4-index trace), respectively.

But a holographic calculation based on the holographic renormalization approach of Skenderis and Papadimitriou (that will not be reproduced here) can be done proving that the coefficients $A(\bar{q})$ and $B(\bar{q})$ are related to the response functions Ω an E at large times as

$$A(\bar{q}) = 4\bar{E}_{(0)}(\bar{q}) \ , \quad B(\bar{q}) = \frac{1}{4}\bar{\Omega}_{(0)}(\bar{q}) \ , \tag{37.27}$$

leading finally to

$$\Delta_S^2(\bar{q}) = \frac{-(i\bar{q})^3}{16\pi^2 \mathrm{Im} B(\bar{q})} \ , \quad \Delta_T^2(\bar{q}) = \frac{-2(i\bar{q})^3}{\pi^2 \mathrm{Im} A(\bar{q})}. \tag{37.28}$$

Note that in principle, the same holographic calculation can be made using the Maldacena map (37.1) to calculate the h_{ij} two-point function from the T_{ij} two-point function on the boundary. The result is the same as above.

The last step is going back to Lorentzian signature, by Wick rotating the momenta, as well as the rank N of the QFT gauge group,

$$\bar{q} = -iq \ , \quad \bar{N} = -iN. \tag{37.29}$$

37.4 Phenomenological QFT Calculation

The quantum field theory in which a calculation is done is the most general super-renormalizable gauge theory on the Euclidean 3 dimensional boundary. In 3 dimensions, Yukawa couplings and ϕ^4 couplings are dimensional, therefore superrenormalizable, and there are no higher powers of fields with dimensional couplings. The fields are gauge field $A_i = A_i^a T_a$, scalars $\phi^M = \phi^{aM} T_a$ and fermions $\psi^L = \psi^{aL} T_a$ which are in the adjoint of $SU(N)$ and have flavor indices M, L.

The Euclidean space action is then

$$\begin{aligned} S_{\mathrm{QFT}} &= \int d^3x \, \mathrm{Tr} \left[\frac{1}{2} F_{ij} F^{ij} + \delta_{M_1 M_2} D_i \Phi^{M_1} \Phi^{M_2} + 2\delta_{L_1 L_2} \bar{\psi}^{L_1} \gamma^i D_i \psi^{L_2} \right. \\ &\quad \left. + \sqrt{2} g_{YM} \mu_{M L_1 L_2} \Phi^M \bar{\psi}^{L_1} \psi^{L_2} + \frac{1}{6} g_{YM}^2 \lambda_{M_1 \dots M_4} \Phi^{M_1} \dots \Phi^{M_4} \right] \\ &= \frac{1}{g_{YM}^2} \int d^3x \, \mathrm{Tr} \left[\frac{1}{2} F_{ij} F^{ij} + \delta_{M_1 M_2} D_i \Phi^{M_1} \Phi^{M_2} + 2\delta_{L_1 L_2} \bar{\psi}^{L_1} \gamma^i D_i \psi^{L_2} \right. \\ &\quad \left. + \sqrt{2} \mu_{M L_1 L_2} \Phi^M \bar{\psi}^{L_1} \psi^{L_2} + \frac{1}{6} \lambda_{M_1 \dots M_4} \Phi^{M_1} \dots \Phi^{M_4} \right] , \end{aligned} \tag{37.30}$$

where in going from the first form to the second we have rescaled the fields by g_{YM}. Here $\lambda_{M_1\ldots M_4}$ and $\mu_{ML_1L_2}$ are dimensionless couplings, $\mathrm{Tr}[T_aT_b] = \frac{1}{2}\delta_{ab}$, and in the first line $[A_i] = 1/2$, $[\Phi^M] = 1/2$, $[\psi^L] = 1$, whereas in the second $[A_i] = 1$, $[\Phi^M] = 1$, $[\psi^L] = 3/2$, as in 4 dimensions.

The fact that in the second form the dimensions are in g_{YM}^2, outside the action, but inside there are no dimensional parameters, is called "generalized conformal structure," and it implies a scaling form for 2-point functions, which in turn implies a scaling form for the coefficients A and B: the dimensions are contained in powers of momenta only, and these come in quantum corrections via unknown functions of the effective coupling

$$g_{\rm eff}^2 = \frac{g^2N}{q}. \tag{37.31}$$

Since also A and B (the whole 2-point function) scales as N^2 in the large N limit, as we saw in general AdS/CFT, we have the general scalings

$$A(q,N) = q^3N^2 f_T(g_{\rm eff}^2)\ , \quad B(q,N) = \frac{1}{4}q^3N^2 f(g_{\rm eff}^2)\ , \tag{37.32}$$

where the 1/4 factor is conventional.

Finally, we need to make the Wick rotation to Lorentzian signature, which takes

$$\bar{q}^3\bar{N}^2 = -iq^3N^2\ , \quad g_{\rm eff}^2(\bar{q},\bar{N}) = g_{\rm eff}^2(q,N)\ , \tag{37.33}$$

and thus the measured superhorizon power spectra in terms of the Euclidean QFT are

$$\Delta_S^2(q) = \frac{q^3}{4\pi^2N^2f(g_{\rm eff}^2)}\ , \quad \Delta_T^2(q) = \frac{2q^3}{\pi^2N^2f_T(g_{\rm eff}^2)}. \tag{37.34}$$

The interesting regime is when the gravitational description is strongly coupled, which results in a weakly coupled dual QFT, meaning we can calculate f and f_T in perturbation theory. Under general considerations, we find in perturbation theory

$$\begin{aligned} f(g_{\rm eff}^2) &= f_0\left[1 - f_1g_{\rm eff}^2\ln g_{\rm eff}^2 + f_2g_{\rm eff}^2 + \mathcal{O}(g_{\rm eff}^4)\right] \\ f_T(g_{\rm eff}^2) &= f_{T0}\left[1 - f_{T1}g_{\rm eff}^2\ln g_{\rm eff}^2 + f_{T2}g_{\rm eff}^2 + \mathcal{O}(g_{\rm eff}^4)\right]. \end{aligned} \tag{37.35}$$

Here f_0 and f_{T0} are obtained from a 1-loop computation, and f_1, f_{T1} and f_2, f_{T2} are from a two-loop computation. In it, we set the RG scale μ equal to the pivot scale q_* of the observational CMBR spectrum.

We can define dimensionless variables g, β and g_T, β_T via

$$\begin{aligned} f_1g_{YM}^2N &= gq_*\ , \quad \ln\beta = -\frac{f_2}{f_1} - \ln|f_1|\ , \\ f_{T1}g_{YM}^2N &= gq_*\ , \quad \ln\beta_T = -\frac{f_{T2}}{f_{T1}} - \ln|f_{T1}|\ , \end{aligned} \tag{37.36}$$

which parametrizes the power spectrum as

$$\Delta_S^2(q) = \frac{\Delta_0^2}{1 + (gq_*/q)\ln|q/\beta g q_*| + \mathcal{O}(gq_*/q)^2}$$
$$\Delta_T^2(q) = \frac{\Delta_{0T}^2}{1 + (g_T q_*/q)\ln|q/\beta_T g q_*| + \mathcal{O}(g_T q_*/q)^2}\,, \tag{37.37}$$

where

$$\Delta_0^2 = \frac{1}{4\pi^2 N^2 f_0}\,, \quad \Delta_{T0}^2 = \frac{2}{\pi^2 N^2 f_{T0}}. \tag{37.38}$$

The parameters f_0, f_1, f_2 (or g, β, Δ_0^2) have been calculated at 2-loops in terms of the number of scalars and fermions, and the dimensionless couplings μ and λ, but the formulas are long and not very illuminating. An important observation is that, in order to be able to fit against CMBR data, we need to introduce a nonminimal coupling between scalars and gravity, in the form of a term in the action of

$$\frac{1}{2g_{YM}^2}\int d^4x\sqrt{-g}\sum_M R(\Phi^M)^2\xi_M. \tag{37.39}$$

As far as the tensor perturbations, the most relevant number is the tensor to scalar ratio, r. In terms of the number of scalars N_S and number of fermions N_F, the formula one finds is

$$r = 32\frac{1 + \sum_{M=1}^{N_S}(1 - 8\xi_M)^2}{1 + 2N_F + N_S}. \tag{37.40}$$

In the limit $N_S \gg N_F$ and $N_S \gg 1$, and for $\xi_M = \xi$ independent of the scalar, we find

$$r \simeq 32(1 - 8\xi)^2, \tag{37.41}$$

The holographic spectrum (37.37) has been compared with the data, as a parametrization, different than the one inspired by inflation with ΛCDM, with an n_s that runs, so n_s and $\alpha_s = dn_s/d\ln q$,

$$\Delta_S^2(q) = \Delta_0^2(q_*)\left(\frac{q}{q_*}\right)^{(n_s-1)+\frac{\alpha_s}{2}\ln\frac{q}{q_*}}. \tag{37.42}$$

The idea in using α_s as a parameter as well is that now we have the same number of parameters: besides the pivot scale q_*, and the normalization $\Delta_0(q_*)$, we have on one side g and β, and on the other n_s and α_s. The result is that both fits (the inflationary one with n_s and α_s, and the phenomenologic holographic cosmology one, with β and g) are compatible with CMBR data.

In the case of the phenomenological holographic cosmology, we find that theories with only gauge fields and fermions are excluded, but if we introduce scalars *and* nonminimal couplings to gravity for them, we can fit the CMBR. To give one example

of a good fit, if we choose a common nonminimality parameter $\xi = 0.133$, giving $r = 0.12$, and a common $\lambda = 1$, we find $N = 2995$ and $N_S = 23255$ (so $N^2 \gg N_S$, as needed).

From the fits, one finds that generically, $f_1 g_{\rm eff}^2 \ln g_{\rm eff}^2 \simeq 1$ at $l \simeq 35$. Thus at very low multipoles ($l < 30$), perturbation theory breaks down, and we would need instead to use lattice quantum field theory to give predictions. Restricting to the reliable part of the data $l > 30$ (known reliably from a theoretical point of view), the holographic cosmology fits the CMBR data just as well as the ΛCDM model with inflation.

37.5 Time Dependent Field Theory Model: Top Down Holography

Another interesting approach to holographic cosmology starts with the original AdS/CFT set-up, of the duality between $\mathcal{N} = 4$ SYM theory in 3+1 dimensions, and string theory (in the supergravity limit) on $AdS_5 \times S^5$, with metric

$$ds_{10}^2 = \frac{R^2}{z^2}[dz^2 + \eta_{\mu\nu} dx^\mu dx^\nu] + R^2 d\Omega_5^2 \,, \tag{37.43}$$

and 5-form

$$F_{(5)} = \omega_5 + *_{10}\omega_5 \,, \tag{37.44}$$

where ω_5 is the volume form on the 5-sphere.

Instead of considering a 4 dimensional bulk cosmology like before, one considers a 5 dimensional bulk, with a spatially flat ($k = 0$) 3+1 dimensional cosmology *near the boundary*, i.e., consider instead a 10 dimensional metric

$$ds_{10}^2 = \frac{R^2}{z^2}[dz^2 + \tilde{g}_{\mu\nu} dx^\mu dx^\nu] + R^2 d\Omega_5^2 \,, \tag{37.45}$$

where the boundary metric is

$$ds_4^2 = \tilde{g}_{\mu\nu} dx^\mu dx^\nu = -dT^2 + a^2(T)\delta_{ij} dx^i dx^j . \tag{37.46}$$

Yet, this is still considered as a *bulk* geometry, with a dual field theory still living in *flat* spacetime, as the boundary is conformally flat:

$$ds_4^2 = a^2(T)\left[-\frac{dT^2}{a^2(T)} + \delta_{ij} dx^i dx^j\right] \equiv a^2(t)[-dt^2 + \delta_{ij} dx^i dx^j] \,, \tag{37.47}$$

where t is conformal time (which we called η in the context of cosmology), which is regular time in the boundary field theory.

In order to be a type of "top-down" model, we need to find a gravitational solution of type IIB that can still be viewed as a dual to some sort of modification of $\mathcal{N} = 4$ SYM. The way this happens is that one introduces a nontrivial dilaton field, specifically time dependent, $\phi = \phi(T)$.

The type IIB supergravity equations of motion with metric and dilaton (Einstein and Klein-Gordon) for our ansatz,

$$\begin{aligned} \tilde{R}_{\mu\nu} &= \frac{1}{2}\partial_\mu\phi\partial_\nu\phi \\ \partial_A(\sqrt{-g}G^{AB}\partial_B\phi) &= 0\,, \end{aligned} \tag{37.48}$$

where g_{AB} is the full 10-dimensional metric and $\tilde{R}_{\mu\nu}$ is the Ricci tensor of $\tilde{g}_{\mu\nu}$, have the solution

$$a(T) = \left(\frac{3|T|}{2R}\right)^{1/3}\,, \quad \phi(T) = \frac{2}{\sqrt{3}}\ln\frac{|T|}{R} + \phi_0\,, \tag{37.49}$$

valid both for $T > 0$ and for $T < 0$. Thus the cosmology presents a Big Crunch, followed by Big Bang, singularity. In conformal time t ($T = \frac{2R}{3}(t/R)^{3/2}$), we have

$$ds_4^2 = \tilde{g}_{\mu\nu}dx^\mu dx^\nu = \frac{|t|}{R}[-dt^2 + \delta_{ij}dx^i dx^j]\,, \quad \phi(t) = \sqrt{3}\ln\frac{|t|}{R} + \phi_0. \tag{37.50}$$

This implies

$$e^\phi = g_s\left(\frac{|t|}{R}\right)^{\sqrt{3}} \tag{37.51}$$

On the boundary theory, since in string theory we have the map

$$\frac{g_{YM}^2}{4\pi} = g_s = e^{\langle\phi\rangle}\,, \tag{37.52}$$

having a time dependent dilaton $\phi = \phi(t)$ amounts to having a time dependent Yang-Mills coupling $g_{YM} = g_{YM}(t)$, specifically

$$g_{YM}^2(t) = g_{YM,0}^2\left(\frac{|t|}{R}\right)^{\sqrt{3}}. \tag{37.53}$$

We thus consider a gauge theory action for the gauge fields

$$S_{YM} = -\frac{1}{4}\int d^4y(g_{YM}^2(t))^{-1}\,\mathrm{Tr}[F_{\mu\nu}F^{\mu\nu}]. \tag{37.54}$$

37.6 Spectrum of Fluctuations Going Through Singularity, from Holography

The specific question to be answered relates to the cosmological evolution near the singularity, where gravity is strongly coupled, therefore not calculable, but the dual field theory is weakly coupled, therefore quite calculable. The question, of relevance for models with Big Crunch followed by Big Bang, is of the propagation of fluctuations through the singularity.

The general strategy is: assume perturbations in the cosmological gravity theory, at $t \to -\infty$, which can be calculated in the gravitational background, or assumed as scale invariant. Translate them to gauge theory, then evolve the gauge theory perturbations, and finally reconstruct the perturbations at $t \to +\infty$.

In the gauge theory, one works in the gauge $A_0 = 0$ and $\partial^i A_i = 0$, so A_i are the only variables. Moreover, one is interested in the IR limit, in which $\vec{k}^2$ is small, so the operator corresponding to the dilaton perturbation $\delta\phi_k$ (in momentum space) is

$$\mathcal{O}_k = \mathrm{Tr}[F^2_{\mu\nu}(k,t)] \simeq 2\dot{A}^2_{i;k}. \tag{37.55}$$

To relate the scalar perturbation to the field theory, we note first that a boundary fluctuation $\delta\phi_k$ must equal a boundary fluctuation of the operator $\mathrm{Tr}[F^2_{\mu\nu}] \simeq 2\dot{A}^2_i$ in momentum space. But

$$\begin{aligned}\dot{A}^2_i &= \int d^3k_1 d^3k_2 \dot{A}_i(k_1)\dot{A}_i(k_2) e^{i(k_1+k_2)x} V \\ &= \int d^3k e^{ikx}\sqrt{V}\int d^3k' \frac{1}{4}\dot{A}_i\left(\frac{k+k'}{2}\right)\dot{A}_i\left(\frac{k-k'}{2}\right)\sqrt{V}\,,\end{aligned} \tag{37.56}$$

where we have defined $k_{1,2} = (k \pm k')/2$. Doing the integral over k, and identifying the Fourier space coefficient with the one of $\delta\phi_k/2$, we obtain

$$M^4_{\mathrm{Pl}}\delta\phi_k = \frac{1}{4}\int d^3k' \dot{A}_i\left(\frac{k+k'}{2}\right)\dot{A}_i\left(\frac{k-k'}{2}\right)\sqrt{V}. \tag{37.57}$$

Then doing the integral in the region $k' < k$, where we can approximately put $k' = 0$ in the arguments of A_i, and obtain $\sim k^3$ from d^3k, we get

$$m^4_{\mathrm{Pl}}\delta\phi_k \sim k^3\dot{A}^2_{i;k}\sqrt{V}. \tag{37.58}$$

If the spectrum of $\delta\phi_k$ is scale invariant, so $\propto k^{-3/2}$ (so $\mathcal{P}_k \sim k^3\delta\phi_k^2$ is k independent), then we obtain the initial gauge perturbation

$$\dot{A}^{\mathrm{in}}_{i;k} \sim k^{-9/4}. \tag{37.59}$$

After performing the gauge theory calculation, we should propagate the operator $\mathcal{O}$ back to the bulk, via a "smearing function", a sort of causal bulk to boundary propagator, that is restricted to tbe nonzero only on a causal wedge relating the point z in the bulk with a slice of time $|t' - t| < z$ on the boundary, by

$$\delta\phi(z, \vec{x}, t) = \int dt' d^3x' K(\vec{x}', t|z, \vec{x}, t)\mathcal{O}(\vec{x}', t'). \tag{37.60}$$

Taking the Fourier transform, we find (defining $\vec{x}' - \vec{x} = \vec{y}'$ and $t' - t = s$)

$$\begin{aligned}\delta\phi(z, k, t) &= V^{-1/2} \int d^3x \delta\phi(z, x, t) e^{ik\cdot x} \\ &= \int ds d^3y' K(\vec{y}', t + s|z, 0, t) e^{-\vec{k}\cdot\vec{y}'} \mathcal{O}(k, t + s) \\ &\simeq \int ds \left[\int d^3y' K(s + t, \vec{y}'|z, 0, t)\right] \mathcal{O}(k, t + s).\end{aligned} \tag{37.61}$$

In this, we must substitute $\mathcal{O}(k, t + s) = 2\dot{A}_i^2$ resulting from the $\dot{A}_i$ in the gauge theory calculation.

This strategy is quite broadly applicable, for various holographic duals.

Results

In the case of the $\mathcal{N} = 4$ SYM with time-dependent coupling, the calculation of the passage through the singularity is as follows. One can redefine fields to

$$\tilde{A}_i = e^{-\phi/2} A_i \ , \tag{37.62}$$

which when substituted back in the gauge theory action gives an equation of motion for a KG field with a time dependent mass,

$$-\partial_\mu \partial^\mu \tilde{A}_i + M^2_{\rm SYM}(t)\tilde{A}_i = 0 \ , \tag{37.63}$$

where

$$M^2_{\rm SYM}(t) = \frac{\ddot{\phi}}{2} - \frac{\dot{\phi}^2}{4} = -\frac{\sqrt{3}(\sqrt{3} + 2)}{4t^2} \ , \tag{37.64}$$

where we can replace $\sqrt{3}$ by whatever power α we have in $g^2_{YM} \propto e^{\phi(t)} \propto |t|^\alpha$. It turns out that, because the mass blows up at the singularity at $t = 0$, there is a singularity in the propagation of the modes through $t = 0$.

To get a sensible result, one needs to regularize, for instance capping off $M_{\rm SYM}(t)$ at $t = -\xi$, maintaining it constant until $t = +\xi$. Then one finds that the spectrum is unchanged as it passes through the singularity (perhaps expected due to the perturbative nature of SYM, but one encounters subtleties along the way, as one has branch cuts). Only the amplitude is changed, between a time $-t$ and a time $+t$ as

$$\tilde{A}_k(+t) \simeq \left(\frac{t}{\xi}\right)^{2\nu_g} \tilde{A}_k(-t) \ , \tag{37.65}$$

where the exponent is

$$2\nu_g = 1 + \sqrt{3} \to 1 + \alpha. \tag{37.66}$$

Then one can go back to the cosmological spectrum, and find that the spectrum as it passes through the perturbation is unchanged, and remains scale invariant. Only the amplitude gets enhanced.

But the calculation is not reliable: we needed to regularize the equations by hand, and even so, we obtain some singular behaviour.

Nevertheless, the method explained here could be more broadly used.

Important Concepts to Remember

- In the holographic cosmology approach, one can equate the wavefunction of the Universe, as a function of spatial metric h_{ij} with the boundary partition function for the same source, $\Psi[h_{ij}] = Z[h_{ij}]$.
- We are interested in correlation functions of h_{ij} in the gravity side, but we can calculate instead correlation functions of the operators dual to them (T_{ij} here) in the dual QFT. We must learn how to relate the two.
- In the domain wall/cosmology correspondence, one can make an Euclidean space continuation (Wick rotation) of a 4 dimensional cosmology with an $a(t)$ to an Euclidean domain wall with $a(z)$, which can have a holographic dual.
- A phenomenological approach to holographic cosmology can be obtained by relating the scalar and tensor power spectra in cosmology, $\Delta^2_{0,S/T}(k)$, with response functions Ω and E that take fluctuations into their canonical conjugates.
- Holographically, Ω and E are calculated from the coefficients A and B of the 2-point functions of T_{ij}, and an analytical continuation to Lorentzian signature.
- As a dual QFT, one considers a general gauge+fermions+scalars+nonminimal coupling to R theory, and matches against CMBR data.
- The matching, considered only over the region $l > 30$ in which the perturbative calculation is valid (as opposed to some lattice QFT one), is as good as ΛCDM with inflation.
- For top-down holography, one can consider a 5 dimensional bulk, deformed $AdS_5 \times S^5$, made up of a 3+1 dimensional flat FLRW metric, which is the metric near the boundary, and is conformal to flat space, and a 4 dimensional $\mathcal{N} = 4$ SYM in flat space.
- One uses a type IIB supergravity solution with a time dependent dilaton and, since $e^{\phi(t)} \sim g_s \sim g^2_{YM}$, a time dependent coupling in $\mathcal{N} = 4$ SYM.
- Fluctuations can be mapped from cosmology to field theory, then evolved through the singularity (which is strongly coupled on the gravity side), then mapped back to cosmology.

Further reading: Maldacena's map holographic map on correlation functions, and from wavefunctions to partition functions, was defined in [70]. The phenomenological holographic cosmology was defined in [71], and its effects for Planck data were studied in [72, 73]. The time dependent field theory model was studied in [74, 75].

Exercises

(1) Write down the 1-loop Feynman diagram(s) for the 2-point function of the energy-momentum tensor in (37.30) and the integral expression using the Feynman rules.
(2) Show that the holographic cosmology ansatz for the scalar spectrum Δ_S^2 in (37.37) and the inflationary ansatz in (37.42) are not that different in terms of their CMBR values, so they have a chance of being alternative fits to the data.
(3) Show that the equations of motion of type IIB supergravity (37.48) have the solutions (37.49).
(4) Show that the equation of motion of the gauge field, under the rescaling $\tilde{A}_i = e^{-\phi/2} A_i$, under the solution (37.49), gives the KG equation with mass (37.64).

References

1. E.W. Kolb and M.S. Turner, "The early Universe," Westview Press, 1994.
2. V. Mukhanov, "Physical foundations of cosmology," Cambridge University Press, 2005.
3. Scott Dodelson, "Modern cosmology," Academic Press, 2003.
4. P.J.E. Peebles, "Principles of Physical Cosmology", Princeton University Press, 1993.
5. F. Couchot, S. Henrot-Versillé, O. Perdereau, S. Plaszczynski, B. Rouillé d'Orfeuil, M. Spinelli and M. Tristram, "Cosmological constraints on the neutrino mass including systematic uncertainties," Astron. Astrophys. **606**, A104 (2017) [arXiv:1703.10829 [astro-ph.CO]].
6. K. Freese, "Review of Observational Evidence for Dark Matter in the Universe and in upcoming searches for Dark Stars," EAS Publ. Ser. **36**, 113 (2009) [arXiv:0812.4005 [astro-ph]].
7. G. D'Amico, M. Kamionkowski and K. Sigurdson, "Dark matter astrophysics," arXiv:0907.1912.
8. D. Clowe, M. Bradac, A.H. Gonzalez, M. Markevitch, S.W. Randall, C. Jones and D. Zaritsky, "A direct empirical proof of the existence of dark matter," Astrophys. J. **648**, L109 (2006) [arXiv:astro-ph/0608407].
9. D.H. Lyth and A.R. Liddle, "The primordial density perturbation: Cosmology, inflation, and the origin of structure," Cambridge University Press, 2009.
10. A.R. Liddle and D.H. Lyth, "Cosmological inflation and large-scale structure," Cambridge University Press, 2000 (for inflation and the CMBR).
11. Daniel Baumann, "TASI lectures on inflation," arXiv:0907.5424 [hep-th].
12. S. Weinberg, "Cosmology", Oxford University Press, 2008.
13. M.J. Duff, B.E.W. Nilsson and C.N. Pope, "Kaluza-Klein Supergravity," Phys. Rept. **130**, 1 (1986).
14. H. Nastase, D. Vaman and P. van Nieuwenhuizen, "Consistency of the AdS(7) x S(4) reduction and the origin of selfduality in odd dimensions," Nucl. Phys. B **581**, 179 (2000) [arXiv:hep-th/9911238].
15. P. West, "Introduction to supersymmetry and supergravity", Wspc, 1990.
16. J. Wess and J. Bagger, "Supersymmetry and supergravity", Princeton University Press, 1992.
17. D. Freedman and A. van Proeyen, "Supergravity", Cambridge University Press, 2012.
18. P. Van Nieuwenhuizen, "Supergravity," Phys. Rept. **68**, 189 (1981).
19. S. Weiberg, "The quantum theory of fields, vol. 3: Supersymmetry", Cambridge University Press, 2005.
20. Barton Zwiebach, "A first course in string theory", Cambridge University Press, 2009.
21. J. Polchinski, "String theory", vols. 1 and 2, Cambridge University Press, 2005.
22. M. Green, J. Schwarz and E. Witten, "Superstring theory", Cambridge University Press, 2012.
23. Clifford V. Johnson, "D-branes", Cambridge University Press, 2003.

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8

24. J.M. Maldacena, "The Large N limit of superconformal field theories and supergravity," Int. J. Theor. Phys. **38**, 1113 (1999) [Adv. Theor. Math. Phys. **2**, 231 (1998)] [arXiv:hep-th/9711200].
25. S.S. Gubser, I.R. Klebanov and A.M. Polyakov, "Gauge theory correlators from noncritical string theory," Phys. Lett. B **428**, 105 (1998) [arXiv:hep-th/9802109].
26. E. Witten, "Anti-de Sitter space and holography," Adv. Theor. Math. Phys. **2**, 253 (1998) [arXiv:hep-th/9802150].
27. Horatiu Nastase, "Introduction to the ADS/CFT Correspondence", Cambridge University Press, 2015.
28. M. Ammon and J. Erdmenger, "Gauge/gravity duality and applications," Cambridge University Press, 2015.
29. D. Baumann and L. McAllister, "Advances in Inflation in String Theory," Ann. Rev. Nucl. Part. Sci. **59**, 67 (2009) [arXiv:0901.0265 [hep-th]].
30. L. McAllister and E. Silverstein, "String Cosmology: A Review," Gen. Rel. Grav. **40**, 565 (2008) [arXiv:0710.2951 [hep-th]].
31. D. Baumann and L. McAllister, "Inflation and String Theory," arXiv:1404.2601 [hep-th].
32. D. Baumann and L. McAllister, "Inflation and String Theory," Cambridge University Press, 2015.
33. H. Bernardo and H. Nastase, "Small field inflation in $\mathcal{N} = 1$ supergravity with a single chiral superfield," JHEP **1609**, 071 (2016) [arXiv:1605.01934 [hep-th]].
34. C.P. Burgess, M. Majumdar, D. Nolte, F. Quevedo, G. Rajesh and R.J. Zhang, "The Inflationary brane anti-brane universe," JHEP **0107**, 047 (2001) [arXiv:hep-th/0105204].
35. H. Nastase and J. Sonnenschein, "Towards brane-antibrane inflation in type IIA: The Holographic MQCD model," JHEP **1202**, 040 (2012) [arXiv:1109.6813 [hep-th]].
36. P. Binetruy, C. Deffayet and D. Langlois, "Nonconventional cosmology from a brane universe," Nucl. Phys. B **565**, 269 (2000) [arXiv:hep-th/9905012].
37. J.M. Cline, C. Grojean and G. Servant, "Cosmological expansion in the presence of extra dimensions," Phys. Rev. Lett. **83**, 4245 (1999) [arXiv:hep-ph/9906523].
38. C. Csaki, M. Graesser, L. Randall and J. Terning, "Cosmology of brane models with radion stabilization," Phys. Rev. D **62**, 045015 (2000) [arXiv:hep-ph/9911406].
39. W.D. Goldberger and M.B. Wise, "Modulus stabilization with bulk fields," Phys. Rev. Lett. **83**, 4922 (1999) [arXiv:hep-ph/9907447].
40. P. Binetruy, C. Deffayet, U. Ellwanger and D. Langlois, "Brane cosmological evolution in a bulk with cosmological constant," Phys. Lett. B **477**, 285 (2000) [arXiv:hep-th/9910219].
41. T. Shiromizu, K.i. Maeda and M. Sasaki, "The Einstein equation on the 3-brane world," Phys. Rev. D **62**, 024012 (2000) [arXiv:gr-qc/9910076].
42. F. Quevedo, "Lectures on string/brane cosmology," Class. Quant. Grav. **19**, 5721 (2002) [arXiv:hep-th/0210292].
43. T. Clifton, P.G. Ferreira, A. Padilla and C. Skordis, "Modified Gravity and Cosmology," Phys. Rept. **513**, 1 (2012) [arXiv:1106.2476 [astro-ph.CO]].
44. N. Arkani-Hamed, S. Dimopoulos and G.R. Dvali, "The Hierarchy problem and new dimensions at a millimeter," Phys. Lett. B **429**, 263 (1998) [arXiv:hep-ph/9803315].
45. I. Antoniadis, N. Arkani-Hamed, S. Dimopoulos and G.R. Dvali, "New dimensions at a millimeter to a Fermi and superstrings at a TeV," Phys. Lett. B **436**, 257 (1998) [arXiv:hep-ph/9804398].
46. L. Randall and R. Sundrum, "A Large mass hierarchy from a small extra dimension," Phys. Rev. Lett. **83**, 3370 (1999) [arXiv:hep-ph/9905221].
47. L. Randall and R. Sundrum, "An Alternative to compactification," Phys. Rev. Lett. **83**, 4690 (1999) [arXiv:hep-th/9906064].
48. D. Marolf and S. Yaida, "Energy conditions and junction conditions," Phys. Rev. D **72**, 044016 (2005) [arXiv:gr-qc/0505048].
49. S. Kachru, R. Kallosh, A.D. Linde and S.P. Trivedi, "De Sitter vacua in string theory," Phys. Rev. D **68**, 046005 (2003) [arXiv:hep-th/0301240].
50. S. Kachru, R. Kallosh, A.D. Linde, J.M. Maldacena, L.P. McAllister and S.P. Trivedi, "Towards inflation in string theory," JCAP **0310**, 013 (2003) [arXiv:hep-th/0308055].

51. H. Abe, T. Higaki and T. Kobayashi, Phys. Rev. D **73**, 046005 (2006) https://doi.org/10.1103/PhysRevD.73.046005 [hep-th/0511160].
52. J. Khoury, B.A. Ovrut, P.J. Steinhardt and N. Turok, "The Ekpyrotic universe: Colliding branes and the origin of the hot big bang," Phys. Rev. D **64**, 123522 (2001) [arXiv:hep-th/0103239].
53. J. Khoury, B.A. Ovrut, N. Seiberg, P.J. Steinhardt and N. Turok, "From big crunch to big bang," Phys. Rev. D **65**, 086007 (2002) [arXiv:hep-th/0108187].
54. J.L. Lehners, "Ekpyrotic and Cyclic Cosmology," Phys. Rept. **465**, 223 (2008) [arXiv:0806.1245 [astro-ph]].
55. P.J. Steinhardt and N. Turok, "Cosmic evolution in a cyclic universe," Phys. Rev. D **65**, 126003 (2002) [arXiv:hep-th/0111098].
56. E.I. Buchbinder, J. Khoury and B.A. Ovrut, "New Ekpyrotic cosmology," Phys. Rev. D **76**, 123503 (2007) [arXiv:hep-th/0702154].
57. R.H. Brandenberger and C. Vafa, "Superstrings in the Early Universe," Nucl. Phys. B **316**, 391 (1989).
58. R.H. Brandenberger, "String Gas Cosmology," String Cosmology, J. Erdmenger (Editor). Wiley, 2009. pp. 193–230 [arXiv:0808.0746 [hep-th]].
59. T. Battefeld and S. Watson, "String gas cosmology," Rev. Mod. Phys. **78**, 435 (2006) [arXiv:hep-th/0510022].
60. R.H. Brandenberger, "String Gas Cosmology after Planck," Class. Quant. Grav. **32**, no. 23, 234002 (2015) [arXiv:1505.02381 [hep-th]].
61. A. Nayeri, R.H. Brandenberger and C. Vafa, "Producing a scale-invariant spectrum of perturbations in a Hagedorn phase of string cosmology," Phys. Rev. Lett. **97**, 021302 (2006) [arXiv:hep-th/0511140].
62. S. Alexander, R.H. Brandenberger and D.A. Easson, "Brane gases in the early universe," Phys. Rev. D **62**, 103509 (2000) [arXiv:hep-th/0005212].
63. J. Khoury and A. Weltman, "Chameleon fields: Awaiting surprises for tests of gravity in space," Phys. Rev. Lett. **93**, 171104 (2004) [arXiv:astro-ph/0309300].
64. J. Khoury and A. Weltman, "Chameleon cosmology," Phys. Rev. D **69**, 044026 (2004) [arXiv:astro-ph/0309411].
65. K. Hinterbichler, J. Khoury and H. Nastase, "Towards a UV Completion for Chameleon Scalar Theories," JHEP **1103**, 061 (2011) Erratum: [JHEP **1106**, 072 (2011)] [arXiv:1012.4462 [hep-th]].
66. L. McAllister, E. Silverstein and A. Westphal, "Gravity Waves and Linear Inflation from Axion Monodromy," Phys. Rev. D **82**, 046003 (2010) [arXiv:0808.0706 [hep-th]].
67. E. Silverstein and A. Westphal, "Monodromy in the CMB: Gravity Waves and String Inflation," Phys. Rev. D **78**, 106003 (2008) [arXiv:0803.3085 [hep-th]].
68. M. Berg, E. Pajer and S. Sjors, "Dante's Inferno," Phys. Rev. D **81**, 103535 (2010) [arXiv:0912.1341 [hep-th]].
69. L. Hui, J.P. Ostriker, S. Tremaine and E. Witten, "Ultralight scalars as cosmological dark matter," Phys. Rev. D **95**, no. 4, 043541 (2017) [arXiv:1610.08297 [astro-ph.CO]].
70. J.M. Maldacena, "Non-Gaussian features of primordial fluctuations in single field inflationary models," JHEP **0305**, 013 (2003) [arXiv:astro-ph/0210603].
71. P. McFadden and K. Skenderis, "Holography for Cosmology," Phys. Rev. D **81**, 021301 (2010) [arXiv:0907.5542 [hep-th]].
72. N. Afshordi, C. Coriano, L. Delle Rose, E. Gould and K. Skenderis, "From Planck data to Planck era: Observational tests of Holographic Cosmology," Phys. Rev. Lett. **118**, no. 4, 041301 (2017) [arXiv:1607.04878 [astro-ph.CO]].
73. N. Afshordi, E. Gould and K. Skenderis, "Constraining holographic cosmology using Planck data," Phys. Rev. D **95**, no. 12, 123505 (2017) [arXiv:1703.05385 [astro-ph.CO]].
74. R.H. Brandenberger, E.G.M. Ferreira, I.A. Morrison, Y.F. Cai, S.R. Das and Y. Wang, "Fluctuations in a cosmology with a spacelike singularity and their gauge theory dual description," Phys. Rev. D **94**, no. 8, 083508 (2016) [arXiv:1601.00231 [hep-th]].

75. E.G.M. Ferreira and R. Brandenberger, "Holographic Curvature Perturbations in a Cosmology with a Space-Like Singularity," JCAP **1607**, no. 07, 030 (2016) [arXiv:1602.08152 [hep-th]].
76. P.A.R. Ade *et al.* [Planck Collaboration], "Planck 2013 results. XXII. Constraints on inflation," Astron. Astrophys. **571**, A22 (2014) [arXiv:1303.5082 [astro-ph.CO]].
77. N. Aghanim *et al.* [Planck Collaboration], "Planck 2018 results. VI. Cosmological parameters," arXiv:1807.06209 [astro-ph.CO].

Index

A
ABJM model, 283
Adiabatic approximation, 293, 381, 384
AdS/CFT correspondence, 271–283, 439
Angular size distance, 29
Anthropic principle, 97
Anti-de Sitter space
  definition, 272
Auxiliary field, 184
Axion, 47, 49, 111, 254, 256, 267, 291, 336, 347, 431
Axion decay constant, 416, 420, 421, 426, 432
Axion inflation, 415
Axion monodromy, 297, 415, 423–426

B
Baryogenesis, 105, 123, 130
Baryon asymmetry problem, 98, 105
Belinfante tensor, 205
Berkovits formulation
  superstring, 238
Betti number, 266, 290, 291
Big Bang Nucleosynthesis, 46, 49, 55, 67, 69–73, 103, 372
BKL singularity, 356
Blackening factor, 312
Black hole, 34
  supermassive, 47
Blue supergiant, 37
Bolometric flux, 31
Boundary condition
  Dirichlet, 207
  Neumann, 207
Brane-antibrane inflation, 314–319
Brane gas, 379, 391–393
Brane probe inflation, 311–314
Braneworld, 263, 311
  D-brane intersections, 268
Braneworld cosmology, 321–329
Breitenlohner–Freedman bound, 281
Brightness, 30
  spectral, 31
Bullet cluster, 43
Bunch-Davies vacuum, 137, 440, 441
Buscher rules, 245

C
Calabi–Yau manifold, 264, 265, 290, 291
Carbon flare-up, 34
Cepheid variable, 33, 37
Chameleon scalar, 269, 292, 405–413
Chan–Paton factor, 231, 234, 251
Charge conjugation matrix, 170
Chern class
  first, 265
Chern number, 415
Chiral multiplet, 184, 299, 301, 332
Cigar geometry, 337, 349
CMBR anisotropy
  dipole, 79
COBE normalization, 78, 145
Cohomology group, 290
Coleman-DiLuccia mechanism
  bubble nucleation, 107
Coma Cluster, 41
Complex structure, 265
Complex structure moduli, 290
Conformal gauge, 208, 213, 385
Conformal group, 276, 280
Conformal invariance, 194, 214, 223, 271, 382, 385

© Springer Nature Switzerland AG 2019

H. Năstase, *Cosmology and String Theory*, Fundamental Theories of Physics 197, https://doi.org/10.1007/978-3-030-15077-8

Conformal Newton gauge, 85, 89
Conformal time
  definition, 23
Consistent truncation, 156, 159, 164–167
Corona, 34
Cosmic frame, 79
Cosmic Microwave Background Radiation (CMBR), 22, 46
  acoustic peaks, 78, 91
  anisotropy, 77
  definition, 4
  kinematics, 79
  spectrum, 80
  temperature, 59
Cosmological constant
  approximate, 113
Cosmological loitering, 388, 392, 393
Cosmological time, 4, 387
Coupling constant
  closed string, 228
  open string, 229
Covariantly constant spinor, 264, 265, 291
Curvature perturbation, 83, 88, 138, 364
Cyclic model, 367–372

D
Dante's Inferno, 426–427
Dark energy, 50, 91, 99, 367, 433
Dark matter, 5, 41, 43
  axion, 47
  candidates, 46
  cold, 47
  hot neutrino, 47
Dark radiation, 328
DBI action, 230, 239, 373, 391
D-brane, 225
  action, 226
Deceleration parameter, 8, 11
Dedekind eta function, 339, 349
de Donder gauge, 173
Density perturbation, 79, 83, 88
Dirac quantization condition, 335
Distance measurements, 34–37
Domain wall/cosmology correspondence, 441
Doppler effect
  relativistic, 3
Doppler shift, 35
D term, 185, 301, 331

E
Effectively massless degrees of freedom
  number of, 58
Einstein static Universe, 273
Ekpyrosis
  two-field, 370
Ekpyrotic phase, 354, 363, 367, 369–371, 375
Ekpyrotic scenario, 353–364
Energy conditions, 12, 21, 361
Entropy
  per baryon, 53, 61, 97, 105
Entropy problem, 97, 105
Equilibrium thermodynamics, 55
Eta problem, 341
  quantum, 296
  string theory, 294, 301, 319
Extrinsic curvature, 322

F
Fast-roll approximation, 369
Fayet-Iliopoulos model, 302
Fayet-Iliopoulos term, 302, 308
Fierz–Pauli action, 173
Fine tuning, 300
Flatness problem, 25, 96, 101, 355
FLRW coordinates, 17, 23
FLRW metric, 15, 29, 271, 362, 434
Freeze-out time, 70
Freund–Rubin ansatz, 188
Friedmann equation, 7, 12, 15, 18, 97, 111
  modified, 328
F term, 184, 301, 331
F-theory, 333
Fuzzy dark matter, 431–437

G
Gauged supergravity, 187
Gauge/gravity duality, 283, 439
Gaugino condensation, 349, 426
Gauss–Codazzi equations, 322, 323
Gauss law
  electromagnetism in d dimensions, 161
  gravity in d dimensions, 162
Generalized conformal structure, 445
Genus
  of surface, 206
G-flux, 267, 292, 333, 345
Ghost condensate, 373, 375
Gibbons-Hawking boundary term for gravity, 323
GKPW construction, 280–282
Global coordinates
  AdS space, 273

Globular clusters, 36, 437
Gluino condensation, 334, 348
Goldberger-Wise stabilization mechanism, 326
Goldstone boson, 50, 203, 416
Graceful exit problem
  old inflation, 107
Grand unified theory, 130
Gravitational lensing, 44
Gravitino, 49, 98, 172, 176, 181, 258, 264, 332
Gravity dual, 283
Gravity modification, 47
Green–Schwarz formulation
  superstring, 237
GSO projection, 237
Gupta–Bleuler quantization, 220
GUT phase transition, 74, 98, 105, 372
GVW superpotential, 292, 333

H
Hagedorn phase, 392, 393
Hagedorn temperature, 379, 397
Halo, 42, 437
Hamilton–Jacobi equation
  for $H(\phi)$, 115
Hertzsprung–Russel diagram, 33, 36
Heterotic superstring, 259, 264
Hidden sector, 48
Hodge number, 291
Holographic cosmology, 439–451
  Maldacena map, 439
Holomorphic 3-form, 266, 290, 333
Holonomy, 264
Homogeneity
  of Universe, 4, 15, 121, 383
Homology, 290
Horizon
  future event, 28
  Hubble, 28
  particle, 28
Horizon distance, 28
Horizon problem, 95, 104, 353, 388, 394
Hot Big Bang, 53
Hubble law, 3, 4

I
Induced metric, 203, 226, 313, 322, 423
Inflation, 120
  eternal, 119
  exponential, 99
  hilltop, 111
  hybrid models, 121
  natural, 111
  new, 99, 111
  old, 99, 106
  paradigm, 98
  power-law, 101
  scalar field, 106
  slow-roll, 111
  Starobinsky, 99
Instanton, 415, 432
Intersection number
  surfaces, 290
Isocurvature perturbation, 83
Isotropy
  of Universe, 4, 15, 121
Israel junction conditions, 322, 326–328, 369

K
Kähler class
  complexified, 291
Kähler form, 265, 291
Kähler manifold, 265
Kähler metric, 185
Kähler potential, 185, 266, 299
Kähler space, 266
Kähler structure moduli, 291
Kähler transformation, 303
Kaluza–Klein (KK)
  background metric, 152, 156
  expansion, 152, 156, 157
  reduction ansatz, 153, 156, 157
Kaluza–Klein compactification, 151, 181
Kaluza–Klein theory
  original, 156
Kappa symmetry, 237
Kibble mechanism, 98
Killing spinor, 183, 186
Killing vector, 183
KKLMMT scenario, 301
  inflation in string theory, 341–350
KKLT model, 331–339, 347, 410–411
Klebanov-Strassler
  geometry, 345
  solution, 337
K-theory, 318

L
ΛCDM model, 41, 43, 50, 431, 446
Large charge approximation, 293
Large extra dimensions, 295, 297, 321, 393
Lepton number, 53

Leptoquark, 130
Level
  string, 221
Level-matching condition, 223
Light-cone gauge, 213, 215
Lightest Supersymmetric Particle (LSP), 48
Lobachevsky space, 272
Local Lorentz symmetry, 155, 174, 265, 387
Luminosity, 31
Luminosity distance, 31
Lyth bound, 144, 297, 427

M
Madelung equations, 435
Magnitude
  absolute, 32
  apparent, 32
Main sequence, 33
Main sequence fitting, 36
Maldacena-Nunez no-go theorem, 331
MAssive Compact Halo Objects (MACHO), 45, 47
Mass measurements, 37
Mathieu equation, 128
Matter domination
  definition, 19
Matter-radiation equality, 61
Measurements of distance by light, 29
Messenger field, 48
Milne Universe, 358
Minimally Supersymmetric Standard Model (MSSM), 48, 296
Mirage cosmology, 322, 329
Mode truncation, 294
Moduli
  complex structure, 266, 269
  Kähler structure, 266, 269
  stabilization, 269
Moduli space
  CY_3, 266
Moduli space approximation, 293, 360, 369
Moduli stabilization
  KKLT Model, 335
  string gas cosmology, 388
MOND, 43, 47, 51
Monopole problem, 98, 105, 372
Moving cluster method, 35
M-theory, 249, 257–259, 283, 333, 353, 354, 359, 363
Mukhanov-Sasaki equation, 138, 140

N
Nambu–Goto action, 202, 226
Navarro-Frank-White profile, 437
Nebula
  planetary, 34
Nepomechie-Teitelboim quantization condition, 336
Neutrino decoupling, 63–65
Neutrino species
  number, 74
Neutron star, 34, 47
New ekpyrotic cosmology, 372–376
Newton constant
  in d dimensions, 163
Newtonian gauge, 134, 138, 394
N-flation, 422
Nijenhuis tensor, 265
Nil 3-manifold, 424
Non-Gaussianity, 428, 439
Nonlinear KK reduction ansatz, 165, 187
Nonrelativistic limit
  scalar field, 433
Nova, 33, 46
NS-NS modes, 254
NSR formulation
  string theory, 237
Null energy condition, 361–364, 372
Number of e-folds, 117

O
Operator/state map, 280
Optical measurements of distance, 30–33

P
Parallax
  spectroscopic, 36
  statistical, 36
  trigonometric, 35
Parallel transport, 264
Parametric resonance, 127
Parsec, 32, 35
Particle
  relativistic, worldline formalism, 191
Particle action
  first order, 195–197
  relativistic second order, 191–194
Peccei–Quinn symmetry, 49, 416
Peculiar velocity, 5, 41, 45, 84
Penrose diagram, 273
Penrose limit, 188
Perfect fluid ansatz, 17
Perturbations problem, 98

Planck constant
in d dimensions, 163
Planck distribution, 22, 77, 95
Planck mass
reduced, 100, 163
Poincaré coordinates
AdS space, 274
Poincaré patch, 273
Polyakov action, 204–205, 207, 209, 219, 223, 382
Pontryagin number, 415
Power spectrum, 137, 140, 143, 364, 400, 428, 446
pp wave, 188
Preheating, 127, 314, 318
instant, 128
Probe approximation, 293
Propagation of light, 27–29
Pullback
on worldsheet, 203
Pure spinor, 237

R
Racetrack
axion, 420
Radiation domination
definition, 20
Radion, 326, 359, 362, 388
Rarita–Schwinger action, 176, 188
Recombination, 61, 62, 67
temperature, 67
Red giant, 33
Redshift
definition, 3
Reheating, 123–129, 134
new, 127
standard, 123
temperature, 127
Relaxation time, 63, 70
Relic, 98, 105, 131, 145
Reparametrization invariance, 192, 203, 205
Ricci-flat, 265
Riemann zeta function, 221
Rotational curves of galaxies, 42, 47
RR Lyrae, 36
R-R modes, 254

S
Sachs-Wolfe
effect, 89
integrated contribution, 90
integrated effect, 78
plateau, 78, 90
Saha equation, 67, 69
Sakharov conditions
baryogenesis, 105, 130
Sasaki-Einstein space, 342
Scale factor duality, 383
Schrödinger equation
in Hubble expansion, 434
S-duality, 256, 259
Sen conjecture
tachyon potential, 318
Silk damping, 88, 91
Slicing
flat, 139
uniform density, 139
Slow roll conditions, 117
Slow-roll inflation, 107, 111–121
Smeared approximation, 293
Smoothness problem, 95, 104, 353
Sound horizon, 87
Spectral tilt, 369
scalar, 142
tensor, 144
Spin connection, 174–178, 265, 267, 359
Standard candle, 13, 27, 36, 37
Static gauge, 208, 226, 313, 391
Stefan–Boltzmann law, 31, 38, 59, 79
Stellar evolution, 33–34
Stiff matter, 20
String duality web, 249, 259
String frame, 254, 362, 393
String gas, 74
power spectra, 399
String gas cosmology, 289, 379–389, 391–402
String tension, 201, 208, 226, 234
String theory
supersymmetric, 237–239, 253
String worldsheet, 201
Sudden decoupling approximation
Cosmic Microwave Background Radiation (CMBR), 89
Superfield
chiral, 184
Superfluid, 434
Supergravity
action, 253
coupled to matter, 186
definition, 172
11 dimensional, 187, 258
first order formulation, 178
minimal in 4 dimensions, 176
second order formulation, 178

Supergravity approximation
string inflation, 299–307
Supergravity models, 181
Supernova, 13, 34, 37, 39, 46
Superpotential, 184, 267, 269, 292, 299–304, 333–336
Supersymmetry, 181
on-shell, 169
rigid, 169
Supersymmetry breaking, 184, 339
gauge mediation, 48, 49
spontaneous, 334
Symmetron, 405, 408

T
Tachyon, 222, 223, 234, 237, 318
Tadpole cancellation, 425
T-duality, 188, 237, 241, 251, 259, 295, 348, 379, 383, 389, 393
closed string, 244–246
open string, 249–253
TeVeS, 44
Thermodynamics
Universe, 58–60
Theta vacuum, 415
Thin-shell effect, 408
't Hooft limit, 280
Torsion, 175
Transfer function
matter, 88

V
Veneziano amplitude, 201, 223
Vielbein, 174, 177, 195
Vielbein postulate, 175
Virasoro constraint, 217, 219, 232, 233, 243
Virial theorem, 41
Visible sector, 48
Visual magnitude
apparent, 4
Volume stabilization
KKLMMT model, 346
Vorticity, 435

W
Wavefunction of the Universe, 440
Weakly Interacting Massive Particles (WIMP), 47, 48
Wess–Zumino model, 169
Wess–Zumino term, 230
Weyl invariance, 205, 206, 213, 382
Weyl tensor, 324
White dwarf, 33, 34, 38, 47
Wilson loop, 251, 268
Winding
string, 242
Worldline formalism
quantum field theory, 194

Y
Yzawa-Yanagida model
supergravity new inflation, 305

Z
Zeta function regularization, 221